.

From Preference Elicitation to Explaining Decisions: a Dialectical Perspective

Habilitation à Diriger des Recherches de l'Université Paris-Saclay

Présentée et soutenue le 8 décembre 2022 par

Wassila Ouerdane

Composition du Jury :

Katie Atkinson Professeure, School of Electrical Enginering, Electronics and Computer Science (EEECS), Université de Liverpool	Rapporteur
Pierre Marquis Professeur, Centre de Recherche en Informatique de Lens (CRIL, CNRS), Univesité d'Artois	Rapporteur
Patrice Perny Professeur, LIP6, CNRS, Sorbonne Unversité	Rapporteur
Madalina Coirtorou Professeure, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Université de Montpellier	Examinateur
Sébastien Destercke Directeur de Recherche, CNRS, Heuristic and Diagnostic of Complex Systems (Heudiasyc,CNRS), Université de Technologie de Compiègne	Examinateur
Nicolas Sabouret Professeur, Laboratoire Interdiciplinaire des Sciences du Numérique (LISN), université Paris-Sacaly	Examinateur
Vincent Mousseau Professeur, Laboratoire Mathématique et Informatique (MICS), CentraleSupélec, Université Paris-Saclay	Parrain HDR

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

This document presents a synthesis of our research work and describes the main results obtained since our PhD [Ouerdane, 2009]. They are the results of numerous and long collaborations with fellow researchers and PhD students.

Our research addresses questions related to knowledge representation and reasoning in the context of eXplainable AI (XAI) [Gunning, 2017]. Our main motivations are designing and modeling adaptive decision support systems to construct and support justified automatic recommendations. Our research lies at the intersection of the fields of Multi-Criteria Decision Aiding (MCDA) and Artificial Intelligence (knowledge representation and reasoning).

Even though we had various opportunities to work on different subjects and domains, the document mainly deals with the various works done within Multi-Criteria Decision Aiding (MCDA) field. Moreover, even if our significant contributions are of the order of formal and theoretical tools, we had several opportunities to be faced with application and real-world contexts with various industrial partners: Decision Brain[1] within the thesis of Lerouge [(in progress)], Dassault Systèmes within the thesis of Tlili [2022], Total within the thesis of Mammeri [2017], IBM within the thesis of El Mernissi [2017], and Place des Leads[2] within the thesis of Maamar [2015]. The focus of the document is mainly on our theoretical contributions. Thus we have not chosen to address these practical aspects and refer the reader to the various PhD thesis for the details.

1.1 Context and Motivations

We are interested in the problems of recommendations, where an "artificial agent adviser" aims to help a user (a decision-maker) build and understand the recommendations for a particular decision problem. Decision aiding is thus a situation involving two parties: a user whose preferences may be incompletely defined or difficult to convey, and an agent, who will have the capabilities to explicitly and accountably represent the reasons for which it recommends a solution to a user [Tsoukiàs, 2008]. Such recommendations mainly stem from Multiple Criteria Decision Aiding models that are well founded from the Decision Theory point of view [Roy, 1996; Bouyssou et al., 2006].

---

[1] https://decisionbrain.com/fr/
[2] Now TimeOne: https://www.timeone.io

Multi-Criteria Decision Aiding (MCDA) aims to develop decision models explicitly based on constructing a set of criteria reflecting the decision-making problem's relevant aspects. These n criteria (often conflicting) ($\mathcal{N} = \{1, 2, \ldots, n\}$ with $n \geq 2$) evaluate a set of alternatives $A = \{a, b, c, ...\}$ from different points of view. Several multi-criteria decision models exist [Bouyssou et al., 2000, 2006]. These models correspond to a parametric family of functions aggregating the evaluation according to each criterion into a solution to the decision problem. The MCDA literature considers different decision problems. We distinguish the *choice*, the *sorting*, the *pairwise comparison*, and the *ranking*. Unlike formulations of choice, ranking and pairwise comparison problems, which are comparative, sorting formulates the decision problem in terms of assigning alternatives to predefined ordered categories $C^1, C^2, ...C^p$, where C^1 (C^p, resp.) is the worst (best, resp.) category. The assignment of an alternative to the appropriate category is based on its intrinsic value and not on its comparison with other alternatives.

In addition, multi-criteria decision aiding results from an interaction between at least two agents, an analyst and a decision-maker. The analyst's goal is to guide the decision-maker (DM) in the construction and understanding of the recommendations of a particular decision problem [Tsoukiàs, 2008]. Decision theory and Multiple Criteria Decision Analysis (MCDA) have established the theoretical foundation upon which many decision support systems have risen. The different approaches (and the formal tools coming along with them) have focused on how a "solution" should be established for a long time. But it is clear that the process involves many other aspects that the analyst handles more or less formally. For instance,

- the problem of accountability of decisions is almost as important as the decision itself. A proper explanation should convince the decision-maker that the proposed solution is the best.

- it should be possible for the decision-maker, to refine, or even contradict, a given recommendation. Indeed, the decision-support process is often constructive because the DM refines its formulation of the problem when confronted with potential solutions.

Let's consider the following situation of decision aiding for illustration. Suppose that a DM wishes to buy a watch. The problem is that once in the store, the person is faced with an extensive choice of models with different colors, sizes, and prices. Impressed and afraid of making mistakes in the selection, he decides to ask for help. Therefore, the seller (referred here by DA for Decision Aider) tries to understand what his customer wants and what are his preferences. After a brief discussion, he notes that from a size point of view, he prefers a small watch to a medium or a big one; he also prefers steel to leather. For the color, he specifies that he likes white more than red or pink and that the watch should be fashion than classical or sport. Finally, the model should be

the less expensive possible. Thus, four models were selected, and their characteristics are depicted in Table 1.1 below.

	Size	Material	Price	Colour	Style
a	small	Steel	450	Red	Classical
b	big	Leather	300	White	Fashion
c	medium	Steel	320	Pink	Classical
d	small	Leather	390	Pink	Sport

Table 1.1: Performance table

On the basis of this information, the DA computes a recommendation and submits it to the DM for a discussion. Such a discussion unfolds as follows:

---

(1) DA: Given your information, b is the best option.

(2) DM: Why is that the case?

(3) DA: Because b is globally better than all other options

(4) DM: What does that mean?

(5) DA: Well... b is top on a majority of criteria considered: the price, the colour, and especially the style, it is so trendy!

(6) DM: But, why b is better than c on the price?

(7) DA: Because c is 20 euros more expensive than b.

(8) DM: I agree, but I see that the guarantee is very expensive especially for this watch. In fact I'm not sure to want the guarantee.

(9) DA : But c remains 5 euros more expensive than b.

(10) DM: I see, but this difference is not significant. And also I changed my mind: I would rather to have a classical model, I think it's more convenient for a daily use.

(11) DA: OK. In this case I recommend c as the best choice.

(12) DM: ...

This made-up scenario involves several aspects that will be discussed in this document.

Let us briefly analyse this dialogue. In turn (1), the DA suggests to the client that b would be the best option for her. The DM challenges this proposition in turn (2) and asks for a justification given by the DA in turn (3). The rationale is based on the fact that the option is better than any other one. Not fully satisfied with this explanation, the DM asks the expert to be more explicit on the reasons motivating his choice. Thus, the DA, in turn (5) explains that b is ranked first on the majority of criteria considered. But, in turn (6), The DM seeks clarification that b is better than another option on a specific criterion. The expert explains that this is since the price of c is more significant than b. We note that this explanation differs from the one given at turn 5. In fact,

unlike turn (4) where the DM wanted to know why b was declared the best choice, in turn (6), he is interested in comparing the model b to another model on a particular criterion. Thus, in turn (5), the DA highlights more explicitly the set of positive points in favour of b regarding the set of all options. In the second case, i.e. turn (6), the DA gave more details on the comparison between two specific models from a particular point of view. Confronted now with such an explanation, the DM rejects it by indicating that the comparison is inappropriate because he doesn't want to include the guarantee in the price. However, in turn (9), the DA maintains that c cannot be better than b because its price is still higher than b. In turn (10), DM indicates that the difference is not significant for her and at the same time, he mentions that he changes her mind about her preferences on the style of the watch. This need to refine or correct old information is very common in practice because a decision-maker is never fully aware of what he wants or prefers at the beginning of the process. Finally, considering the DM's remarks, the DA suggests that, now, c is a better choice.

This example dialogue illustrates how different types of explanations can be asked (and provided) and how the available information may change and be corrected (because the decision-maker really changes his mind, but also because the expert necessarily makes some assumptions that only hold by default). This is especially true when the decision-maker is confronted with explicit justifications because it helps him to identify relevant questions and possible critics.

1.2 Research Questions and Contributions

Our objective is to design artificial agents able to serve as analysts (like in the previous example within a recommender system context, for instance) for various meaningful decision-aiding contexts, capable of initiating and steering a dialogue with a user to derive a recommendation, alternating between the elicitation of preference information, and the presentation of complete or partial recommendations. Prompted by the user, an agent should support its assertions with explanations and would gently steer the conversation towards the production of a recommendation which is fully agreed upon, potentially following a non-monotonic path in its representation of the user's preference - reconsidering pieces of information or even the preference model in the light of the user's responses. Communication with the user should be simple but faithful to the rich information conveyed and in line with the context of the decision-aiding situation. In other terms, we aim to handle and take into account the different aspects of a decision-aiding process by adopting the perspective of *an interactive approach* whereby:

- Preference elicitation can be done incrementally, taking into account the feedback of the user (such as contradicting a previous assertion, asking for an explanation,

etc.) to fit the user's model as well as possible while minimizing at the same time the cognitive effort of the user; and

- Justification (or explanation) can be given to the user on the proposed items or on facts inferred by the adviser during the interaction so that the user can correct or contradict the relevant information.

Such an interactive approach requires a sufficiently expressive means to convey the agent's messages. It is important to note that in our research work, the communication between the agent and the user will not rely on advanced techniques of natural language processing, which is, on the other hand, an open door for new research and future collaborations (see Chapter 5). Instead, the interaction will be guided by a structured dialogue, designed as a set of rules regulating the interaction [Walton and Krabbe, 1995; Carlson, 1983; Ferguson et al., 1996; McBurney and Parsons, 2003]. Thus, the communication with the adviser will happen through a set of possible utterances chosen by the user.

We structured our research lines around two main topics to reach our objectives.

1.2.1 Modeling and generating explanations for recommendations for complex decision problems.

The question of explaining a decision, recommendation, algorithm outputs, etc., often associated in the literature with the acronym XAI (eXplainable AI) [Gunning, 2017; Barredo Arrieta et al., 2020], has become in recent years a crucial element in any "trusted algorithmic design". Indeed, for high-stakes AI applications, performance is not the only criterion to consider. Such applications may require a relative understanding of the logic executed by the system. In this case, the end-user wants an answer to the question "Why?". eXplainable Artificial Intelligence (XAI) aims to provide methods that help empower AIs to answer this question. Even though interest in this question has exploded with machine learning tools and techniques [Biran and Cotton, 2017; Gilpin et al., 2018; Guidotti et al., 2019; Mohseni et al., 2018; Barredo Arrieta et al., 2020], it dates back to expert systems [Swartout, 1983; Gregor and Benbasat, 1999], and since then, many works have emerged. Various questions are explored, such as: generating and providing explanations, identifying desirable characteristics of an explanation from the point of view of its recipient, evaluating the explanation produced by the system, etc. [Herlocker et al., 2000; Carenini and Moore, 2006; Tintarev, 2007; Nunes et al., 2014; Doshi-Velez and Kim, 2017; Miller, 2019; Vilone and Longo, 2021]

Our work focuses on *designing and implementing tools and algorithms for generating explanations for recommendations stemming from multi-criteria models* which put user *preferences and judgments* at the heart of the reasoning. Generating explanations in the

MCDA context is not a simple task; as different criteria are at stake, the user cannot fully assess their importance or understand how they interact. Moreover, once the user is faced with the result and the explanation, he may realize that it is not exactly what he expected. Therefore, it can make changes or provide new information that will have effects, for example, on the other phases of the decision-aiding process (e.g., the preferences learning step). Thus, beyond making the result acceptable, presenting an explanation can impact the representation of the user's reasoning mode, which is at the base of the construction of the recommendation. Furthermore, the challenge with this question is that the concept of explanation varies depending on the decision context/problem and the decision model. Indeed, as the requirements vary significantly from situation to situation (for instance, depending on the criticality of the stakes and the time pressure) and from decision-maker to decision-maker, we do not believe in providing a unique explanation. Indeed, our approach stems from a set of patterns for different types of explanation (depending on the decision model under use and the user's profile), allowing tailored answers to the user. Under such perspectives, our research work intends to answer the following question:

> *Given a decision model and a set of preference information, is there a principled way to define a simple complete explanation supporting a recommendation/decision?*

To answer the previous question, we addressed mainly two MCDA decision models[3]: one very widely used model, whether in decision theory or machine learning, namely the *additive model* and the other which is the *Non-Compensatory Sorting (NCS)* model [Bouyssou and Marchant, 2007a,b]. With the first model, the different contributions aimed to explore the concept of explanations for pairwise comparisons (why is one option better than another?) or choice problems (why an option is the best?). In contrast, in the second, we seek to explain the assignment of an alternative to a given category (why is an option classified in the worst category? for instance). The following Table 1.2 gathers all our contributions for this topic, and the details are given in Chapter 4.

Decision Problem	Model	References
Choice	Weighted Majority	[Labreuche et al., 2011]
	Additive Utility	[Labreuche et al., 2012]
Pairwises Comparisons	Additive Utility	[Belahcene et al., 2019], [Belahcene et al., 2017a]
Sorting	NCS	[Belahcene et al., 2017b], [Belahcene et al., 2018b]

Table 1.2: Our Contributions to the Explainability Topic for MCDA

---

[3]We were also interested in other models/systems, for example, rule-based systems (classical and fuzzy) and optimization models, which are not detailed in this document. We refer the reader to [El Mernissi, 2017; Baaj, 2022; Lerouge, (in progress)] for more details.

Our proposals are based on different approaches and techniques: argument schemes [Walton, 1996] and mathematical programming. In particular, the question of constructing explanations comes down to formalizing argument schemes that link premises (information provided or approved by the user or deduced during the process of preference learning, and some additional hypotheses on the process of reasoning (from the assumptions of the model) to a conclusion (e.g. the recommendation). By casting the reasoning steps under the form of argument schemes, we make explicit assumptions usually hidden for the decision-maker, hence allowing meaningful explanations.

Finally, in all of our works on constructing and designing explanations, we seek to follow (when it is possible) some key principles of explanations (see *e.g.* [Miller, 2019; Coste-Marquis and Marquis, 2020]):

- Explanation shall be rigorous (important decision) ⤳ One shall bring proof (complete explanation)

- Explanation shall be understandable ⤳ One shall define a language which relates directly to the preferential information (e.g. not include the weights). In other words, we want explanations to be conveyed in an expressive language to the recipient of this explanation.

- Explanation shall be relevant ⤳ One shall define what could be pertinent to focus on within the decision situation. For instance, mentioning neutral elements (that do not influence the decision) may seem irrelevant and should be avoided if possible.

- Explanation shall be simple ⤳ One shall define different levels of complexity. We want explanations to be "easy to process" by the recipient of the explanation.

1.2.2 Modeling the interaction for constructing adaptive decision support systems.

At present, when decision-aiding support or recommendation systems (online, for example) are in full expansion, an important aspect is that of succeeding in capturing and integrating the preferences, habits, and reactions of users to try to produce the most compelling and relevant recommendations from a user perspective. To meet this objective, we investigated two lines of research.

Setting up efficient preference learning and elicitation mechanisms : Learning and eliciting preferences is essential in a decision support process. This step aims to incorporate user judgments (preferences) as faithfully as possible into the decision model. Developing relevant and reliable recommendations is crucial, and any flawed process would lead to unsubstantiated advice being provided to users. In addition,

preferences are essential in many contexts, such as decision-making, machine learning, recommendation systems, social choice theory, and various sub-fields of Artificial Intelligence (see, for instance, [Jacquet-Lagrèze and Siskos, 2001; Peintner et al., 2008; Kaci, 2011; Furnkranz and Hullermeier, 2011; Hüllermeier, 2014; Pigozzi et al., 2016]). In this context, the challenge is to build learning algorithms that are both efficient (from a computational point of view) while keeping humans in the loop to integrate and represent their expertise and skills knowledge as faithfully as possible.

The basic idea of the multi-criteria decision support methodology is that, given a decision problem, we collect preferential information from the DM to build an evaluation model. This model must reflect the point of view (the value system) of the DM and help him to solve the decision problem. In other words, our research is interested in implementing efficient algorithms to learn models' parameters using the information contained in reference examples–a training set. This is what we call (*indirect elicitation* or *learning from examples*). In this context, we follow an (indirect) approach, close to a machine learning paradigm [Furnkranz and Hullermeier, 2011], where a set of reference assignments is given and assumed to describe the decision-maker's point of view. The aim is to *extend* these assignments with this decision model. Thus, we sought to answer the following question:

> *For a given decision situation, assuming that a given decision model is relevant to structure the decision maker's preferences, what should be the parameters' values to fully specify this model that corresponds to the decision-maker viewpoint?*

To answer this question, we worked on different models: the Non-Compensatory Sorting model, its variant the MR-Sort model [Leroy et al., 2011] and the Ranking with Multiple Profiles (RMP) method [Rolland, 2013]. The different contributions are summarized in Table 1.3 below. The different proposals seek to offer tools that, on the one hand, will provide more efficient devices (in terms of computation time), and on the other hand, extend the literature to consider new types of preferential information. More precisely, we rely on logical formalism (Boolean-based) to meet the first need. Second, we investigate the question of building preference learning tools in the case of non-monotone preferences (single-peaked [Black, 1958]).

Designing adaptive dialectical system   We are interested in a decision-aiding process (as illustrated in Section 1.1). In this context, there are at least two distinct actors: a decision-maker (DM), and an analyst, whom we shall call in what follows a decision aider (DA). Both play very different roles [Tsoukiàs, 2007]. The DM has some preferences on the decision options and is, in the end, responsible for the decision to be taken and justifying it. The DA helps him in this task by bringing some methodology

Methods		Approaches	
		MIP-based	Boolean-based
Sorting	NCS	[Leroy et al., 2011]	[Belahcene et al., 2018a] [Tlili et al., 2022]
	MR-sort	[Minoungou et al., 2020], [Minoungou et al., 2022]	
Ranking	RMP	[Liu et al., 2014], [Olteanu et al., 2021]	[Belahcene et al., 2018c]

Table 1.3: Our Contributions to preference Learning & Elicitation Topic

and rationality. The DA analyses the consistency of the information provided by the DM, proposes some recommendation based on such information and construct the corresponding justifications. A key ingredient of the decision process is how interaction takes place. In particular, the DA should be able to adapt to the DM's responses. In fact, the DM's preferences are often incomplete or not fixed at the beginning of the process. Only when confronted with the recommendation and its justification the DM can react and give relevant feedback. The competence of a human DA is precisely to integrate this new information, to revise his representation of the profile of the DM so as to produce a finely adapted recommendation that can be understood and accepted.

Now, there are many different contexts in which decision aiding can take place, and an artificial agent sometimes plays the role of the DA. Take, for instance, recommender systems used on commercial websites: the role of the DA is to suggest items that the DM is likely to buy (travel, books, etc.). Often the product space is vast, and the DA's role is to help navigate this catalog. According to [McGinty and Smyth, 2006], "user feedback is a vital component of most recommenders". Moreover, to take this feedback into account timely and consistently, some authors argue to maintain a preference model of the user [Viappiani et al., 2006]. Model-based recommendation systems are then based on a unique model (e.g. the additive utility) and rely upon the assumption that all potential users can be represented by this model [Viappiani et al., 2006]. However, in the case of multi-criteria recommendation, there is a wide variety of possible preference models, and assuming a fixed model may prove too restrictive. In other terms, rather than making an assumption that may later be found to be incorrect (as an example: the weighted mean model is often used in many systems but without an explicit justification), our idea is to simultaneously reason with several possible models and let the system decide the one appropriate to the current user. With this assumption, our research work seeks to answer the following question:

> *How to equip an artificial agent with adaptive behavior and model the system's reasoning to allow "efficient" interaction with a user within a decision-aiding situation?*

Setting up such an automatic system to support this interaction raises several questions. If the agent can choose among several models, is there a principled way to do so? Would such a method be dependent on the models considered? How do we make a formal link between the generation of the explanation and the improvement of the preference learning process? Indeed, faced with an explanation, a user can provide new information, invalidate old one etc. These reactions strongly contribute to feeding the learning phase of the preference model. How to adapt classic preference learning algorithms to manage inconsistent user feedback (inconsistency, erroneous information, etc.) while automatically adjusting the model to the information provided by the user?

Our research aims to provide a formal language to represent such an interaction, explain it, communicate its results, and convince the user that what is happening is theoretically sound and operationally reasonable. Most of the work in this direction has been initiated within our PhD [Ouerdane, 2009], and the different contributions are summarized in the following Table 1.4.

Approach	References
Argumentation-based interaction	[Ouerdane et al., 2011] [Ouerdane, 2009] [Ouerdane et al., 2010] [Ouerdane et al., 2008] [Labreuche et al., 2015]

Table 1.4: Our Contributions to the Interaction Topic

In these contributions, we concentrated on some questions : (i) if the DA can choose among several models, is there a principled way to do so? (ii) would such a method be dependent of the models considered? And, finally (iii) how, in practice, should such an interaction be regulated?

We borrow from decision theory and Multiple Criteria Decision Analysis to answer the first point in the positive. Regarding (ii), we advocate a generic method to account for this adaptive behavior. Indeed, instead of focusing on a given collection of models, we adopt an axiomatic approach, and thus characterize which models can be handled in the way we propose. As for (iii), the actual procedure we put forward takes the form of a dialogue game between the DM and the DA, and is inspired by recent work in dialectical management and dialogue systems resulting from work in multi-agent systems and argumentation theory [McBurney and Parsons, 2003; Black et al., 2021]. We proposed to build and formalize an interaction protocol, which specifies the rules and conditions under which we can have a "coherent" interaction in a decision support context where the initiative is sometimes left to the user (e.g. ask for an explanation). The details are given in Chapter 5.

The other issues, as we shall see in Chapter 5, are a rich source of future works and collaborations.

1.3 Structure and Content of the Document

- **Chapter 2: MCDA: Concepts and Definitions** is devoted to describing the Multiple Criteria Decision Aiding concepts used in the different contributions. We will restrict ourselves to addressing only the necessary materials for the following chapters. More precisely, we describe the components of a preference elicitation process. Moreover, we present two aggregation methods: the additive model and the Non-Compensatory Sorting model. Indeed, our different contributions are mainly related to these two models.

- **Chapter 3: Efficient Tools for Preference Learning and Elicitation** exposes the different mathematical and computational tools implemented to address the question of learning the parameters of the NCS model and its variants (U^B-NCS: a unique profile, U^C-NCS: a unique set of sufficient coalitions and MR-Sort: additive coalitions). Concretely, we proposed two formulations based on Boolean satisfiability to learn the parameters of the Non-Compensatory Sorting model from perfect preference information, i.e. when the set of reference assignments can be wholly represented in the model. We also extend the two formulations to handle inconsistency in the preference information by adopting the Maximum Satisfiability problem language (MaxSAT). These formulations are described in the first part of the chapter. The second one extends the literature to consider new types of preferential information for learning the parameters of the MR-Sort model, such as the fact that preferences on criteria are not necessarily monotone but possibly single-peaked (or single-valley) [Black, 1948, 1958].

- **Chapter 4: Supporting Decisions: a panel of explainability tools** addresses our developments of explainability tools within the MCDA context. In this context, our main concern is developing principle-based approaches and cognitively bounded models of explanations. By principle-based approach, we mean that each explanation is attached to a number of well-understood properties of the underlying decision model. By cognitively bounded, we suggest that the statements composed of an explanation will be constrained to remain easy to grasp by the receiver (decision-maker). We investigated different decision models (Additive utility, NCS) and various decision problems (Choice, pairwise comparisons and sorting). In our proposal, we rely on numerous tools from AI (argument schemes [Walton, 1996]) and mathematical programming to formalize and compute explanations and their contents.

- **Chapter 5: Interactive recommendations and explanations.** is devoted to discussing the dialectical perspective that we want to set up to formalize the interaction between an artificial agent adviser and a user. In this interaction, elicitation, recommendation and explanation are tightly interleaved. In the first

part of the chapter, we present our preliminary works in this direction. The second part describes all the perspectives and the mid and long-term research works that we plan to have in the following years with different collaborations.

The document is based on a collection of papers available in Appendix C. Many of these works have also been conducted in the context of some PhD co-supervision. Specifically, designing efficient algorithms for preference elicitation, described in Chapter 3, have been studied in the PhD of Jinyan Liu (co-supervised with Vincent Mousseau, MICS, CentraleSupélec), Pegdewedé Stéphane Minoungou (co-supervised with Vincent Mousseau and Paolo Scotton, IBM Zurich) and Ali Tlili (co-supervised with Vincent Mousseau and Oumaima Khaled, Dassault Systèmes). The question of constructing explanations for MCDA addressed in Chapter 4 was the central question studied in the PhD of Khaled Belahcene (co-supervised with Vincent Mousseau, Nicolas Maudet – Lip6, Sorbonne univeristé) and Christophe Labreuche –Thales). Finally, Manuel Amoussou started last year a PhD on this topic by taking this interaction perspective (co-supervised with Vincent Mousseau and in collaboration with Nicolas Maudet and Khaled Belahcene, Heudiasyc, Université de Technologie de Compiegne) .

MCDA: Concepts and Definitions

We devote this chapter to describing and defining the different concepts in Multi-Criteria Decision Aiding (MCDA) used in our various contributions. We will restrict ourselves to addressing only the necessary materials for the following chapters. We do not intend to do a literature review as the present document is dedicated only to summarize our research work.

2.1 Multiple Criteria Decision Aiding

Decision aiding results from an interaction between an "analyst" (or expert) and a "client" (or decision-maker – DM). The analyst aims to guide the decision-maker to find a solution to his problem and to be convinced that this solution is a good one [Tsoukiàs, 2008; Bouyssou et al., 2006]. Within this context, MCDA is an umbrella term to describe a collection of formal approaches which seek to take explicit account of multiple criteria (points of view) in helping individuals or groups explore decisions that matter. More formally, MCDA accounts for $\mathcal{N} = \{1, 2, \ldots, n\}$ *points of view* (criteria) evaluating a set of *alternatives* $\mathbb{X} = \{x, y, z, \ldots\}$.

We assume the points of view provide a sense of the relative performance of alternatives, for which two representations could be considered:

- *preference profiles*, a tuple $\langle \succsim_i \rangle_{i \in \mathcal{N}} \in (\mathbb{X} \times \mathbb{X})^{\mathcal{N}}$ of *total preorders* over alternatives – binary relations that are transitive. This representation is often used in Social Choice or when representing preferences with an outranking relation[1]. Example 2.1 provides an illustration with a situation detailed in Chapter 4 where each point of view corresponds to the views of a juror in a jury $\mathcal{N} = \{ \unicode{x2640}^1, \unicode{x2642}^2, \unicode{x2640}^3, \unicode{x2642}^4, \unicode{x2640}^5 \}$ gathered to assess the performance of a number of candidates $\{a, b, c, d, e, f\} \subseteq \mathbb{X}$. Each preference profile details the ordinal preferences of jurors over candidates. Here we have total orders - there are no ties.

- *performance tables*, where an alternative $x \in \mathbb{X}$ is described by a tuple of performance scalars $\langle x_i \rangle_{i \in \mathcal{N}}$ encoding its performance according to each point of

---

[1]An outranking relation naturally provides four outcomes when comparing two alternatives: preference for the former, for the latter, indifference, or incomparability; also, it does not enforce transitivity of preference [Bouyssou, 2009; Roy, 1991]

view $i \in \mathcal{N}$ on an ordinal scale $(K_i, \geq_i)$. Table 2.1 provides an illustration with alternatives representing cars, situation used to illustrate the functioning of an aggregation model, see Example 2.3 in this chapter.

Example 2.1̃(Example of preference profiles)

$\text{⚥}^1:$ $a \succ_1 b \succ_1 f \succ_1 e \succ_1 c \succ_1 d$

$\text{⚥}^2:$ $e \succ_2 b \succ_2 c \succ_2 d \succ_2 a \succ_2 f$

$\text{⚥}^3:$ $f \succ_3 a \succ_3 b \succ_3 d \succ_3 e \succ_3 c$

$\text{⚥}^4:$ $d \succ_4 a \succ_4 c \succ_4 e \succ_4 f \succ_4 b$

$\text{⚥}^5:$ $c \succ_5 e \succ_5 b \succ_5 f \succ_5 d \succ_5 a$

Example 2.2̃(Example of a performance table)

Alternatives m_i are car models, described according to cost, acceleration, braking and road holding. Cost is measured in dollars, acceleration is measured by the time, in seconds, to reach 100 km/h from full stop—lower is better, braking power and road holding are both measured on a qualitative scale ranging from 1 (lowest performance) to 4 (best performance).

car model	cost	acceleration	braking	road holding
m_1	16 973	29	2.66	2.5
m_2	18 342	30.7	2.33	3
m_3	15 335	30.2	2	2.5
m_4	18 971	28	2.33	2
m_5	17 537	28.3	2.33	2.75
m_6	15 131	29.7	1.66	1.75

Table 2.1: A performance table for car model evaluation

The basic idea in decision aiding methodology is that, given a decision problem, we collect preferential information from the decision-maker such that his system of values is either faithfully represented or critically constructed, in order to build a model which, when applied, should turn a recommendation for action to the decision-maker. Under such a perspective, a fundamental step is acquiring preferential information from a decision-maker, or as it is commonly named preference learning and elicitation process [Furnkranz and Hullermeier, 2011].

2.2 Preference Learning and Elicitation Process

Preferences are fundamental to decision processes since the recommendations are meaningful and acceptable only if the decision-makers' values are considered. Within this context, a challenging activity is "preference learning and elicitation", which aims to capture the DMs' preferences to specify the decision model parameters accurately. The challenge is related to the nature of the preferences expressed by the DMs, which can be imprecise, conflicting, unstable, time-dependent, yet they should be structured and synthesized. This elicitation process can be implemented in many ways. In this section, we give a high-level description of it and quickly review its components.

2.2.1 A brief description

The different components of the elicitation process are depicted in Figure 2.1.
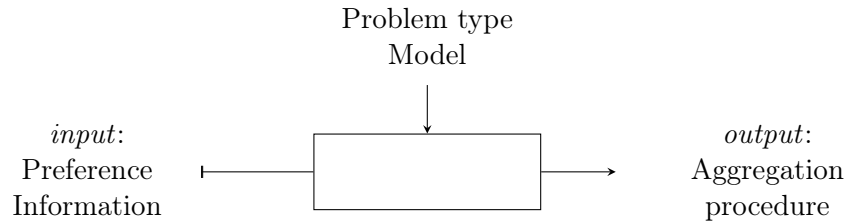
Figure 2.1: The elicitation process.

Preference information. It encompasses any information provided by the decision-maker to the learning process. The following questions concerning preference information organize the elicitation process:

1. What type of preference information should be obtained?

2. How to collect preference information?

3. How preference information should be processed so as to sculpt the aggregation procedure?

4. How to account for imperfect preference information?

All these questions need to be considered carefully, and there are many different ways to address each one.

Type of problem.   Different decision problems exist. They are represented in Figure 2.2:

- *sorting* problems consist in assigning alternatives to categories, known in advance and ordered by level of requirement;

- *pairwise comparison* problems consist in deciding, for each pair of alternatives, which one is the better;

- *choice* problems consist in selecting the "best" alternative or a subset of "best" alternatives among any group;

- *ranking* problems consist in ordering the group of options from the worst to the best, with possible ties.

$\star$	$\star\star$	$\star\star\star$
ac	bdf	e

$\{b \succ a, d \succ c\}$ d $e \succ f \succ d \succ c$
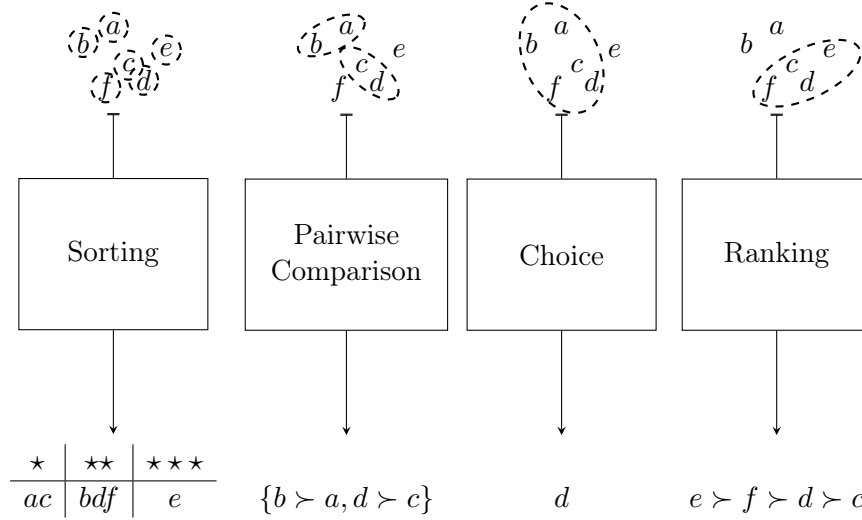
Figure 2.2: Aggregation procedures.

We note that the points of view, the way the alternatives are described according to each point of view, and the type of problem are contextual elements that need to be provided to the elicitation process. They are usually defined in a preliminary phase, called *problem structuring* [Bouyssou et al., 2000], which is out of the scope of this work.

Aggregation procedures.   The elicitation process is expected to output an *aggregation procedure*, whose role is to bring together several (conflicting) points of view into a single overall judgment. More precisely, the aim is to obtain an aggregation procedure that: i) reflects the views of the decision-maker and ii) helps him solve his decision problem.

2.2.2 The aggregation model

Technically, an aggregation model consists of a parameterized family of aggregation procedures. Each value of the *preference parameter* specifies a single aggregation procedure. For instance, in a weighted sum the preference parameter are the weights corresponding to the importance of the different criteria involved in the decision problem. Therefore, the goal of the elicitation process is to interpret the preference information to pinpoint the values of the preference parameters to yield the corresponding procedure. Moreover, the aggregation models can be sorted into three families [Perny, 2000; Grabisch and Labreuche, 2010; Rolland, 2013]:

- *Aggregate, then compare*: the approach aims at computing an overall numeric score, the *value* for each alternative, representing the overall performance of an alternative. Then, the usual ordering of numbers is used to compare alternatives. An example of a method following this approach is the one of the additive model (see Section 2.3.1).

- *Compare, then aggregate*: In this approach the preferences according to each point of view need to be synthesized into an *outranking relation* denoting overall preference. Then, this relation is *exploited* to yield an answer permitting to sort, choose or rank alternatives (e.g. NCS and MR-Sort methods, see Section 2.3.2).

- *Rule-based systems*: Monotonic rules, of the form 'if an alternative is at least/at most as good as such alternative according to such point of view, then . . . ' have been used to formally describe preferences for a long time (e.g. *expert systems* [Waterman, 1986] implementing decision trees). This type of aggregation will not be discussed in this manuscript.

Moreover, a critical step (decision) in an elicitation process is to select a model. The selection of which approach to use in a specific decision making context is not a trivial one, and this choice needs to be based on the particular characteristics of the problem under analysis (see for guidelines [Guitouni and Martel, 1998; Bouyssou et al., 2000; Roy and Słowiński, 2013]). This question of choosing/selecting a model is not the mainstream of the work described in this document. Still, as we shall see in Chapter 5, we believe that this question can be tightly related to the provision of an explanation to the decision-maker within the decision-aiding process.

2.2.3 How to specify an aggregation model?

When a model has been chosen, one issue is to assess the model's parameters. One way, referred to as *elicitation* (or direct elicitation), requires the participation of the DM, whose preferences and values have to be incorporated into the model. Elicitation

proceeds by asking questions to the DM to set the required parameter values. Note that by "direct elicitation", we do not mean questioning the model's parameters values directly. It has been abundantly argued in the literature (see [Podinovskii, 1994; Roy and Mousseau, 1996],Bouyssou et al. [2006, §4.4.1]) that questioning, for instance, about importance of criteria weighted is bad practice.

Another way is known as *learning* (or *indirect elicitation*, or *disaggregation paradigm* [Doumpos and Zopounidis, 2011]). The model parameters are inferred based on reference examples (for instance, in sorting problem, we have assignment examples). This approach is close to the machine learning paradigm [2]. In this approach, preference information is considered as external *data*, and the elicitation process has to do with an input that is limited in length and quality but hopefully meaningful. The idea is to transform *holistic preferences* information into information about the parameters governing the aggregation procedure.

Finally, in a decision-aiding process, the availability of DMs is usually limited. Therefore, it is important to ask the DM informative questions. This is what is called "Active Learning" [Benabbou et al., 2017; Kadziński and Ciomek, 2021]. In this setting, a "budget of questions" is available. They should be chosen adequately, either in sequence or all from the start. Appropriate criteria for selecting questions have to be studied.

In our work related to building efficient algorithms for learning preferences (see Chapter 3), we adopted the second approach. In our setting, holistic preferences take the form of either pairwise, ordinal preference statements such as alternative 'a is preferred to alternative b', when considering a pairwise comparison problem, or the assignment of some alternative to some category, when considering a sorting problem (see Figure 2.2). Hence, in the first phase, preference statements about alternatives are translated into statements about parameters; then, we may face different situations, that is, either the *set of parameters compatible* with these statements is:

- *Empty.* Therefore, either the analyst decides to extend the aggregation model, or he tries to find the parameters' values that 'best reflect' the statements of the decision-maker by asking more questions; or

- *Reduced to a singleton.* In this situation, the elicitation is complete (the corresponding model matches the point of view of the decision-maker); or

- *Larger* (contains more than one element). Thus, either more preference information is collected, or specific values of the preference parameters are singled out from the set of values compatible with the preference information[3].

---

[2]The interested reader may want to see the interesting review paper by [Doumpos and Zopounidis, 2011]

[3]Many methods exist to implement a choice function yielding 'the most representative preference

2.3 Focus on Some Aggregation Models

In our various contributions, we have considered two families of models: additive models (aggregate and compare paradigm) and outranking models (compare then aggregate paradigm). In what follows, we describe the two models on which we constructed our various contributions.

2.3.1 Additive utility model

A preference relation $\succsim$ follows a *value model* when a numerical score can measure the overall desirability of an alternative; the higher, the better. Technically, there is a numeric function $\mathcal{U}$ mapping alternatives to real numbers:

$$\mathcal{U}: \quad \mathbb{X} \quad \longrightarrow \quad \mathbb{R}$$
$$x = (x_1, \ldots, x_n) \longmapsto \sum_{i=1}^{n} u_i(x_i)$$

Scores are then compared to derive preferences:

$$\forall x, y \in \mathbb{X}, \ x \succsim y \iff \mathcal{U}(x) \geq \mathcal{U}(y) \tag{2.1}$$

This way of comparing alternatives produces a preference relation that is both *transitive* —i.e. for any alternatives $x, y, z \in \mathbb{X}$, if $x \succsim y$ and $y \succsim z$, then $x \succsim z$— and *complete*—i.e. for any alternatives $x, y \in \mathbb{X}$, either $x \succsim y$, or $y \succsim x$, or both—in which case we say x is *indifferent* or *equally preferred* to y, and we denote $x \sim y$. Reciprocally, any binary relation that is transitive and complete can be represented in the value model, without too much loss of generality.

In MCDA, the role of the additive value model is central. It is the flagship of value models—those described in the *aggregate then compare* paradigm (see Section 2.2). It serves as the basis of very popular methods, such as the *multi-attribute value theory* (MAVT) [Keeney and Raiffa, 1976]. It is also used in Machine Learning. Classifiers are functions that map objects, often described by tuples of features, to categories. If the features can be interpreted as measuring some desirability, this behavior can be considered through the prism of the aggregation of evaluations stemming from multiple points of view.

2.3.2 Non-Compensatory Sorting model

Multi-criteria sorting aims at assigning alternatives to one of the predefined ordered categories $C^1 \prec \ldots \prec C^p$. All alternatives are evaluated on n criteria, $\mathcal{N} = \{1, 2, \ldots, n\}$; hence, an alternative a is characterized by its evaluation vector $(a_1, \ldots, a_n)$, with $a_i \in \mathbb{X}_i$

---

parameters', hence, the 'most representative aggregation procedure'. For more details, we refer the reader, for instance, to [Kadzinski et al., 2012; Siskos et al., 2005; Furnkranz and Hullermeier, 2011].

denoting its evaluation on criterion i. Each criterion is equipped with a weak preference relation $\succsim_i$ defined on $\mathbb{X}_i$. We assume, without loss of generality, that the preference on each criterion increases with the evaluation (the greater, the better). We denote by $\mathbb{X} = \prod_{i \in \mathcal{N}} \mathbb{X}_i$ the Cartesian product of evaluation scales.

We recall in what follows the definitions of an upset and the upper closure of a subset w.r.t. a binary relation:

Definition 2.1 (Upset and upper closure). *Let $\mathcal{A}$ be a set and $\mathcal{R}$ a binary relation on $\mathcal{A}$.*

- *An upset of $(\mathcal{A}, \mathcal{R})$ is a subset $\mathcal{B} \subseteq \mathcal{A}$ such that $\forall a \in \mathcal{A}, \forall b \in \mathcal{B}, a\mathcal{R}b \Rightarrow a \in \mathcal{B}$.*

- *The upper closure $cl_{\mathcal{A}}^{\mathcal{R}}(\mathcal{B})$ of a subset $\mathcal{B} \subseteq \mathcal{A}$ is the smallest upset of $(\mathcal{A}, \mathcal{R})$ containing it. : $\forall \mathcal{B} \subseteq \mathcal{A}, cl_{\mathcal{A}}^{\mathcal{R}}(\mathcal{B}) := \{a \in \mathcal{A} : \exists b \in \mathcal{B}\ a\mathcal{R}b\}$.*

Non-Compensatory Sorting (NCS) method [Bouyssou and Marchant, 2007a,b] is a MCDA sorting model originating from the ELECTRE TRI method [Roy, 1991]. NCS can be intuitively formulated as follows: an alternative is assigned to a category if: *i)* it is better than the lower limit of the category on a sufficiently strong subset of criteria, and *ii)* this is not the case when comparing the alternative to the upper limit of the category.

In what follows, we introduce NCS formally considering the case of two categories and the one with multiple categories.

2.3.2.1   Sorting into two categories

In the Non-Compensatory Sorting model (NCS), limiting profiles defines the boundaries between categories. Therefore, a single profile corresponds to the case where alternatives are sorted between two ordered categories that we label as GOOD and BAD. A pair of parameters describes a specific sorting procedure:

- a limiting profile $b \equiv \langle b_i \rangle_{i \in \mathcal{N}}$ that defines, according to each criterion $i \in \mathcal{N}$, an upper set $\mathcal{A}_i \subset \mathbb{X}_i$ of approved values at least as good as b_i (and, by contrast, a lower set $\mathbb{X} \setminus \mathcal{A}_i \subset \mathbb{X}_i$ of disapproved values strictly worse than b_i), and

- a set $\mathcal{T}$ of sufficient coalitions of criteria, which satisfies monotonicity with respect to inclusion.

These notions are combined into the following assignment rule:

$$\forall x \in \mathbb{X}, \quad x \in \text{GOOD} \iff \{i \in \mathcal{N} : x_i \succsim_i b_i\} \in \mathcal{T} \tag{2.2}$$

An alternative is considered as GOOD if, and only if, it is better than the limiting profile b according to a sufficient coalition of criteria. By considering the approved sets, the rule can be equivalently written as follows:

$$\forall x \in \mathbb{X}, \quad x \in \text{GOOD} \iff \{i \in \mathcal{N} : x_i \in \mathcal{A}_i\} \in \mathcal{T} \tag{2.3}$$

2.3.2.2 Sorting into multiple categories

With p categories, the parameter space is extended accordingly, with approved sets $\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, \, k \in [2..p]}$ defined by a set of limiting profiles $\langle b_i^k \rangle_{i \in \mathcal{N}, \, k \in [2..p]}$ and sufficient coalitions $\langle \mathcal{T}^k \rangle_{k \in [2..p]}$ declined per boundary. The ordering of the categories $\{C^1 \prec \ldots \prec C^p\}$ translates into a nesting of the sufficient coalitions: $\forall k \in [2..p]$, $\mathcal{T}^k$ is an upset of $(2^{\mathcal{N}}, \subseteq$) and $\mathcal{T}^2 \supseteq \cdots \supseteq \mathcal{T}^p$, and also a nesting of the approved sets: $\forall i \in \mathcal{N}, \forall k \in [2..p]$, $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \precsim_i)$ and $\mathcal{A}_i^2 \supseteq \cdots \supseteq \mathcal{A}_i^p$. These tuples of parameters are augmented on both ends with trivial values: $\mathcal{T}^1 = \mathcal{P}(\mathcal{N})$, $\mathcal{T}^{p+1} = \emptyset$, and $\forall i \in \mathcal{N}$, $\mathcal{A}_i^2 = \mathbb{X}$, $\mathcal{A}_i^{p+1} = \emptyset$.

With $\omega = (\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, \, k \in [2..p]}, \langle \mathcal{T}^k \rangle_{k \in [2..p]})$, Bouyssou and Marchant [2007b] define the sorting function NCS_ω from $\mathbb{X}$ to $\{C^1 \prec \ldots \prec C^p\}$ with the following rule:

$$NCS_\omega(x) = C^k \Leftrightarrow \begin{cases} \forall k' \leq k, & \{i \in \mathcal{N} : x \in \mathcal{A}_i^{k'}\} \in \mathcal{T}^{k'} \text{ and} \\ \forall k' > k, & \{i \in \mathcal{N} : x \in \mathcal{A}_i^{k'}\} \notin \mathcal{T}^{k'}. \end{cases} \tag{2.4}$$

Note that Bouyssou and Marchant [2007a,b] define a broader class of sorting method which includes vetoes: it is possible for a single criterion to forbid the assignment to a category. Throughout this document, we only consider NCS without veto; therefore, we should formally write NCS without veto all along with the document. However, to facilitate the reading, we choose to write NCS even if we consider NCS model without a veto.

Example 2.3 illustrates the functioning of the NCS model. It summarizes how we aggregate the preference information to get an overall assignment of the different car models. Before applying such a model, we need to set up through an elicitation process the limiting profiles and the sufficient coalitions of criteria.

Example 2.3. An illustrative example for NCS

A journalist prepares a car review for a forthcoming issue. He considers a number of popular car models and wants to sort them to present a sample of cars "selected for you by the editorial board" to the readers. This selection is based on four criteria: cost (€), acceleration (time, in seconds, to reach 100 km/h from full stop – lower is better), braking power and road holding, both measured on a qualitative scale ranging from 1 (lowest performance) to 4 (best performance). The performances of the six models are described in Table 2.2.

model	cost	acceleration	braking	road holding
m_1	16 973€	29.0 sec.	2.66	2.5
m_2	18 342€	30.7 sec.	2.33	3
m_3	15 335€	30.2 sec.	2	2.5
m_4	18 971€	28.0 sec.	2.33	2
m_5	17 537€	28.3 sec.	2.33	2.75
m_6	15 131€	29.7 sec.	1.66	1.75

Table 2.2: Performance table for models of cars.

In order to assign these models to a category among $C^{1^\star}$ (average) $\prec C^{2^\star}$ (good) $\prec C^{3^\star}$ (excellent), the journalist considers an NCS model:

- The attributes of each model are sorted between average ($\star$/ ■), good ($\star\star$/ ■) and excellent ($\star\star\star$/ ■) by comparison to the profiles given in Table 2.3.

Profile	cost	acceleration	braking	road holding
$b^{1^\star}$	17 250€	30.0 sec.	2.2	1.9
$b^{2^\star}$	15 500€	28.8 sec.	2.5	2.6

Table 2.3: Limiting profiles.

The resulting labeling of the six alternatives according to each criterion is depicted in Figure 2.3 and Table 2.4.
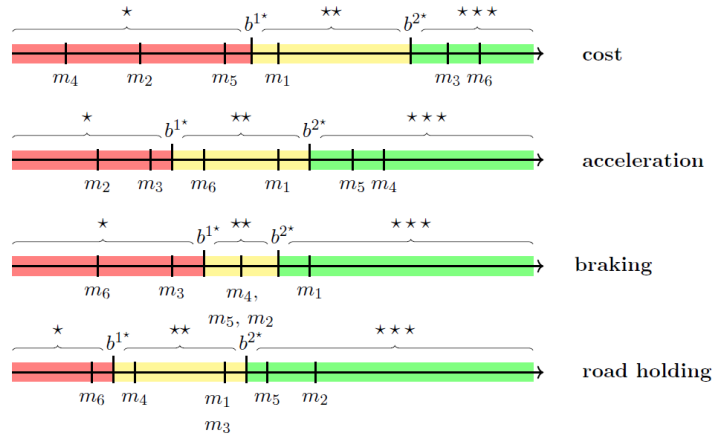
Figure 2.3: Representation of performances w.r.t. category limits.

model	cost	acceleration	braking	road holding
m_1	★★	★★	★★★	★★
m_2	★	★	★★	★★★
m_3	★★★	★	★	★★
m_4	★	★★★	★★	★★
m_5	★	★★★	★★	★★★
m_6	★★★	★★	★	★

Table 2.4: Categorization of performances.

- These appreciations are then aggregated by the following rule: *an alternative is categorized good or excellent if it is good or excellent on cost or acceleration, and good or excellent on braking or road holding. It is categorized excellent if it is excellent on cost or acceleration, and excellent on braking or road holding.* Being excellent on some criterion does not really help to be considered good overall, as expected from a Non-Compensatory model. Sufficient coalitions are represented on Figure 2.4 (where arrows denote coalition strength). Finally, the model yields the assignment presented in Table 2.5.

Alternatives	m_1	m_2	m_3	m_4	m_5	m_6
Assignment	★★	★	★★	★★	★★★	★

Table 2.5: Alternative assignments.
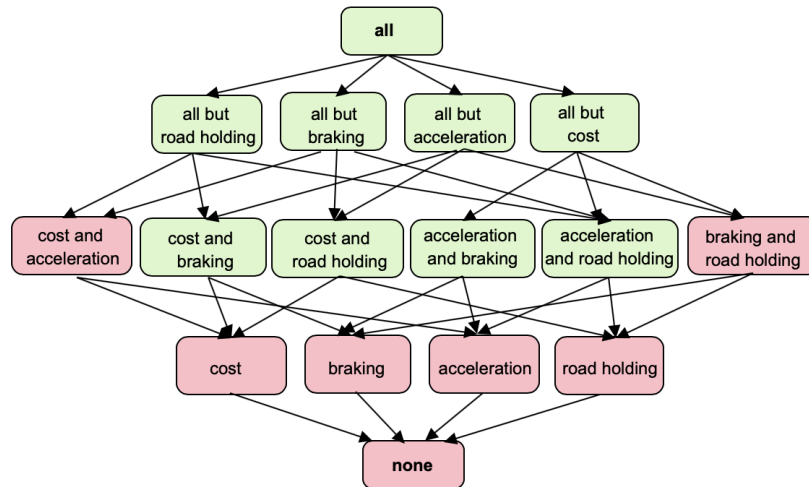
Figure 2.4: Sufficient (green) and insufficient (red) coalitions of criteria

2.3.2.3   Variants of the NCS Model

A number of variants of the Non-Compensatory Sorting model can be found in the literature. On the one hand, as it was mentioned previously, Bouyssou and Marchant [2007a,b] define the NCS classes of sorting methods, which includes the possibility of vetoes. On the other hand, there exist variants, without veto, corresponding to simplifications of the model, with additional assumptions that restrict the parameters—limiting profiles and sufficient coalitions—either explicitly or implicitly.

Following Bouyssou and Marchant [2007b], one may consider to explicitly restrict either the sequence of limiting profiles, or the sequence of sufficient coalitions:

- U^C-NCS: Non-Compensatory Sorting with a unique set of sufficient coalitions: $\mathcal{T}^2 = \cdots = \mathcal{T}^p$;

- U^B-NCS: Non-Compensatory Sorting with a unique boundary/limiting profile $b^2 = \cdots = b^p$ or, equivalently, $\forall i \in \mathcal{N},\ \mathcal{A}_i^2 = \cdots = \mathcal{A}_i^p$.

It is worth noting that an NCS model which is in U^C-NCS and U^B-NCS simultaneously corresponds necessarily to a model with two categories.

A particular case of NCS corresponds to Majority Rule Sorting (MR-Sort) model [Leroy et al., 2011]: when the families of sufficient coalitions are all equal $\mathcal{F}^2 = ... = \mathcal{F}^p = \mathcal{F}$ and defined using additive weights attached to criteria, and a threshold: $\mathcal{F} = \{F \subseteq \mathcal{N} : \sum_{i \in F} w_i \geq \lambda\}$, with $w_i \geq 0$, $\sum_i w_i = 1$, and $\lambda \in [0,1]$. Moreover, as the finite set of possible values on criterion i, $\mathbb{X}_i = [min_i, max_i] \subset \mathbb{R}$, the order on $\mathbb{R}$ induces a complete pre-order $\succcurlyeq_i$ on $\mathbb{X}_i$. Hence, the sets of approved values on criterion i, $\mathcal{A}_i^h \subseteq \mathbb{X}_i$ ($i \in \mathcal{N}, h = 2...p$) are defined by $\succcurlyeq_i$ and $b_i^h \in \mathbb{X}_i$ the minimal approved value in $\mathbb{X}_i$ at level h: $\mathcal{A}_i^h = \{x_i \in \mathbb{X}_i : x_i \succcurlyeq_i b_i^h\}$. In this way, $b^h = (b_1^h, \ldots, b_n^h)$ is interpreted as the frontier between categories C^{h-1} and C^h; $b^1 = (min_1, ..., min_n)$ and $b^{p+1} = (max_1, ..., max_n)$ are the lower frontier of C^1 and the upper frontier of C^p, respectively. Therefore, the MR-Sort rule can be expressed as:

$$x \in C^h \quad \text{iff} \quad \sum_{i:x_i \geq b_i^h} w_i \geq \lambda \text{ and } \sum_{i:x_i \geq b_i^{h+1}} w_i < \lambda \qquad (2.5)$$

It should be emphasized that in the above definition of the MR-Sort rule, the approved sets $\mathcal{A}_i^h$ can be defined using $b^h \in \mathbb{X}$, which are interpreted as frontiers between consecutive categories, only if preferences $\succcurlyeq_i$ on criterion i are supposed to be *monotone*. Thus, a criterion can be either defined as a *gain* or a *cost* criterion:

Definition 2.2. *A criterion $i \in \mathcal{N}$ is:*

- *a gain criterion: when $x_i \geq x_i' \Rightarrow x_i \succcurlyeq_i x_i'$*

- *a cost criterion: when* $x_i \leq x'_i \Rightarrow x_i \succcurlyeq_i x'_i$

Therefore, in case of:

- a gain criterion, we have $x_i \in \mathcal{A}_i^h$ and $x'_i \geq x_i \Rightarrow x'_i \in \mathcal{A}_i^h$, and $x_i \notin \mathcal{A}_i^h$ and $x_i > x'_i \Rightarrow x'_i \notin \mathcal{A}_i^h$. Therefore $\mathcal{A}_i^h$ is specified by $b_i^h \in \mathbb{X}_i$: $\mathcal{A}_i^h = \{x_i \in \mathbb{X}_i : x_i \geq b_i^h\}$.

- a cost criterion, we have $x_i \in \mathcal{A}_i^h$ and $x'_i \leq x_i \Rightarrow x'_i \in \mathcal{A}_i^h$, and $x_i \notin \mathcal{A}_i^h$ and $x_i < x'_i \Rightarrow x'_i \notin \mathcal{A}_i^h$. Therefore $\mathcal{A}_i^h$ is specified by $b_i \in \mathbb{X}_i$: $\mathcal{A}_i^h = \{x_i \in \mathbb{X}_i : x_i \leq b_i^h\}$.

We shall see in the next chapter how we can adapt these definitions to consider new kinds of preference information. More specifically, we were interested in extending the literature for preference elicitation to non-monotone data.

2.4 Summary

This chapter introduces the different notations and concepts we shall use in the following chapters. As discussed initially, an essential step in the decision-aiding process is the preference elicitation process. This activity aims to make the decision maker's preferences explicit through a model representing them. In other terms, it consists of determining plausible values (or ranges of variation) for the parameters of the chosen model based on the preference information provided by the decision-maker. To do so, it is necessary to design efficient procedures and algorithms to specify this model and its parameters. In Chapter 3 we summarized our contributions to this aim, by considering NCS and MR-Sort models.

Efficient Tools for Preference Learning and Elicitation

3.1   Introduction

The subject of "preferences" has gained considerable attention in Artificial Intelligence. It has become a new interdisciplinary research area closely linked to related fields such as operations research, social choice theory, and decision theory [Ozturk et al., 2005; Kaci, 2011; Furnkranz and Hullermeier, 2011]. It is about constructing methods to learn preference models from implicit or explicit preferences, which are used to capture, model and predict the preferences of an individual or group of individuals.

Under such a perspective, our work is situated within the Multi-Criteria Decision Aiding field, where there is a need to structure the decision-aiding process in which a decision-maker (DM) and an analyst interact to build a multi-criteria preference model. The expected advantage of this process is to provide insights into the decision problem and lead to recommendations regarding the decision to be made. Within the decision-aiding process, the process by which the analyst and the DM interact is called an elicitation process. This process aims to incorporate the DM's judgments into the preference model. Within this context, our works contribute to providing formal tools for the following question:

> "For a given decision situation, assuming that a given decision model is relevant to structure the decision maker's preferences, what should be the parameters' values to fully specify the model that corresponds to the decision-maker viewpoint?"

To address this issue, we have carried out several works, with a significant part dedicated to the Non-Compensatory Sorting (NCS) model and its variants: U^B-NCS, U^C-NCS and MR-Sort (see Chapter 2). In this chapter, we trace the landscape, summarized in Table 3.1, of the different mathematical and computational tools that we have implemented to address the question of learning the parameters of the NCS model (and its variants).

Methods		Approaches	
		MIP-based	Boolean-based
Sorting	NCS	[Leroy et al., 2011]	[Belahcene et al., 2018a] [Tlili et al., 2022]
	MR-sort	[Minoungou et al., 2020], [Minoungou et al., 2022]	
Ranking	RMP	[Liu et al., 2014], [Olteanu et al., 2021]	[Belahcène et al., 2023]

Table 3.1: Contributions to preference learning and elicitation

The different proposals seek to offer tools that, on the one hand, will provide more efficient devices (in terms of computation time) by appealing to logical formalism—on the other hand, extend the literature to consider new types of preferential information, such as the fact that preferences on criteria are not necessarily monotone but possibly single-peaked [Black, 1948, 1958]. Moreover, the set of tools has an important theoretical significance. Still, it can also serve as a base for practical applications–see, e.g. [Belahcene et al., 2018b] for an application in an *accountability* setting (see Chapter 4 for more details). Finally, in addition to sorting models, we also proposed tools for learning the parameters of the Ranking with Multiple Profiles Method (RMP) [Rolland, 2013]. This work is briefly described at the end of this document. We refer the interested reader to [Liu et al., 2014; Olteanu et al., 2021; Belahcene et al., 2018c] for more details.

3.2   Learning NCS Model Parameters

The Non-Compensatory Sorting model aims to assign alternatives evaluated on multiple criteria to one of the predefined ordered categories (see Chapter 2). Two popular variants of the NCS model are the NCS model with a unique profile (U^B-NCS) and the NCS model with a unique set of sufficient coalitions (UC-NCS). Moreover, another variant of NCS is the one in which the importance of criteria is additively represented using weights: the MR-Sort model (see Chapter 2).

Before exposing our contributions, let us recall the problems of learning the parameters of the NCS model and its variant MR-Sort, named Inv-NCS and Inv-MR Sort problems, respectively.

The Inv-NCS problem   We define the inverse Non-Compensatory Sorting problem as a decision problem, where the input is some preference information under the form of an ordinal performance table concerning a set of reference alternatives and an assignment of these reference alternatives to categories (see Example 2.3), that gives a

positive answer if, and only if, there is a preference parameter of the Non-Compensatory Sorting model (i.e. a tuple of approved sets and a tuple of approved coalitions satisfying some monotonicity constraints), which is consistent with this preference information. Formally,

An *instance* of the Inv-NCS problem is a sextuple $(\mathcal{N}, \mathbb{X}, \langle \succsim_i \rangle_{i \in \mathcal{N}}, \mathbb{X}^\star, \{C^1 \prec \ldots \prec C^p\}, \alpha)$ where:

- $\mathcal{N}$ is a set of criteria;

- $\mathbb{X}$ is a set of *alternatives*;

- $\langle \succsim_i \rangle_{i \in \mathcal{N}} \in \mathbb{X}^2$ are *preferences* on criterion i, $i \in \mathcal{N}$, $\succsim_i \subset \mathbb{X}^2$ is a total pre-ordering of alternatives according to this criterion;

- $\mathbb{X}^\star \subset \mathbb{X}$ is a finite set of *reference alternatives*;

- $\{C^1 \prec \ldots \prec C^p\}$ is a finite set of *categories* totally ordered by *exigence level*.

- $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$ is an *assignment* of the reference alternatives to the categories. Therefore, 'α^{-1}' is the associated inverse function i.e. for a given category C^h, $\alpha^{-1}(C^h) = \{x \in \mathbb{X}^\star : x \in C^h\}$.

When referring to an instance, we shorten this sextuple as 'α'. Thus, a *solution* of the instance α of the Inv-NCS problem is a parameter $\omega = (\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, \, k \in [2..p]}, \langle \mathcal{T}^k \rangle_{k \in [2..p]})$ of the NCS model (see Section 2.3.2) such that $\forall x \in \mathbb{X}^\star, \; \alpha(x) = NCS_\omega(x)$.

The Inv-MR-Sort problem Considering as input a learning set L, which is the couple $(A^*, \mathcal{C})$, where $\mathcal{C} = \{cat(a), \forall a \in A^*\}$; that is each alternative $a \in A^* \subset \mathbb{X}$ is assigned to a desired category $cat(a) \in \{1, \ldots, p\}$. Therefore, the Inv-MR-Sort problem consists in taking as input this learning set L and computes the parameters of the MR-Sort method, namely the weights (w), the majority level (λ) and the limit profiles (b), that best restore L, i.e. maximizing the number of correct assignments.

3.3 SAT/MaxSAT Formulations for Inv-NCS

For learning the parameters of an NCS model, we follow an (indirect) approach, close to a machine learning paradigm [Furnkranz and Hullermeier, 2011], where a set of reference assignments is given and assumed to describe the decision-maker's point of view. The aim is to *extend* these assignments with an NCS model (see Section 2.2.3). We have shown in [Belahcene et al., 2018b] that Inv-NCS problem is NP-Hard

Until now, indirect approaches to the elicitation of Non-Compensatory Sorting models based on mathematical programming ([Leroy et al., 2011]) suffer from poor computational efficiency, that restrict them to solving toy instances. To cope with the computation burden, a heuristic approach has been proposed [Sobrie et al., 2015, 2019] which can handle large datasets, but lose optimality guaranty. To cope with the computation burden without losing optimality guarantee, we investigated a novel direction based on Boolean satifiability formulation (SAT). In short, a Boolean satisfaction problem consists in a set of Boolean variables V and a logical proposition about these variables $f : \{0,1\}^V \rightarrow \{0,1\}$. A solution $v^\star$ is an assignment of the variables mapped to 1 by the proposition: $f(v^\star) = 1$. A binary satisfaction problem for which there exists at least one solution is *satisfiable*, else it is *unsatisfiable*. Without loss of generality, the proposition f can be assumed to be written in conjunctive normal form: $f = \bigwedge_{c \in \mathcal{C}} c$, where each *clause* $c \in \mathcal{C}$ is itself a disjunction of literals, which are variables or their negation $\forall c \in \mathcal{C}, \exists c^+, c^- \in \mathcal{P}(V) : c = \bigvee_{v \in c^+} v \vee \bigvee_{v \in c^-} \neg v$, so that a solution satisfies at least one condition (either positive or negative) of every clause.

Concretely, we proposed two formulations based on Boolean satisfiability to learn the parameters of the Non-Compensatory Sorting model from perfect preference information, i.e. when the set of reference assignments can be wholly represented in the model. We also extend the two formulations to handle inconsistency in the preference information by adopting the Maximum Satisfiability problem language (MaxSAT). We start by summarizing the contribution in the case of perfect preference information.

3.3.1   SAT-based formulations for Inv-NCS

Hereafter, we summarize two formulations of the Inv-NCS problem in the framework of Boolean satisfiability. The idea is to reduce the problem of finding the parameters of an NCS model faithfully reproducing a given assignment of alternatives to categories to the SAT problem of finding an assignment of Boolean variables that verifies a given propositional formula written in conjunctive normal form.

We proposed two formulas stem from different representation strategies. One, described in Section 3.3.1.1, establishes a bijection between the parameter space of the NCS model and the valuation of the propositional variables. The second detailed in Section 3.3.1.2 leverages a powerful representation theorem that allows keeping implicit the set of coalitions by introducing the notion of *pairwise separation* using pairs of alternatives given in the assignment..

In other terms, when using the representation strategy based on the explicit representation of the set of coalitions of criteria, each solution of the SAT/MaxSAT problem found by the solver can directly be interpreted in terms of parameters of an NCS model (either of the U^B or the U^C subtype). This is not precisely the case with the representa-

tion strategy based on pairwise separation of alternatives: the SAT/MaxSAT solution explicitly describes the approved sets of value on each criterion and at each satisfaction level (i.e. the boundary profiles), but the sets of sufficient coalitions are left implicit. They are solely described in terms of an upper and a lower bound.

3.3.1.1   SAT formulation based on Coalitions

A first formulation Φ_α^C was introduced in [Belahcene et al., 2018a; Belahcene, 2018]. It is based on an explicit representation of the parameter space of the NCS model – coalitions of points of view $\langle\ \mathcal{T}^k\ \rangle$ and approved sets of alternatives $\langle\ \mathcal{A}_i^k\ \rangle$, for each point of view $i \in \mathcal{N}$ and each level of exigence $k \in [2..p]$ – leading to a formulation in conjunctive normal form with $\mathcal{O}(2^{|\mathcal{N}|} + p \times |\mathcal{N}| \times |\mathbb{X}^\star|)$ variables and $\mathcal{O}(p \times |\mathbb{X}^\star| \times 2^{|\mathcal{N}|})$ clauses, such that $\mathcal{N}$ is the set of criteria, $\mathbb{X}^\star$ is the set of assignment examples and p the number of categories.

We provide here an informal presentation of the approach; formal justification can be found in [Belahcene et al., 2018a; Tlili et al., 2022]. The explicit representation Φ_α^C involves two families of binary variables.

- The first family (denoted a) defines the approved sets according to the set of criteria such that for given alternative, level and criterion, the associated variable equals 1 if and only if the alternative is approved at the considered level according to the considered criterion.

- The second family (denoted t) of binary variables uniquely specifies the set of sufficient coalitions for each level i.e. given a coalition of criteria, the associated variable equals 1 if and only if the coalition is sufficient.

The SAT formulation *based on coalitions* aims at learning both NCS parameters $(\langle\ \mathcal{A}_i^k\ \rangle_{i\in\mathcal{N},\ k\in[2..p]}, \langle\ \mathcal{T}^k\ \rangle_{k\in[2..p]})$ from a set of assignment examples, thus, two types of clauses are considered. The first type of clauses (ϕ_α^{Ci}, $i \in [1..4]$, below) defines these parameters and reproduces the structural conditions i.e.: the monotonocity of scales, approved sets and sufficient coalitions sets are ordered by inclusion. The second type of clauses (ϕ_α^{C5} and ϕ_α^{C6}, below) ensures the restoration of the assignment examples.

Clauses.   For a Boolean function written in conjunctive normal form, the clauses are *constraints* that must be satisfied simultaneously by any antecedent of 1. The formulation Φ_α^C is built using six types of clauses:

- Clauses ϕ_α^{C1} ensure that each approved set $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}^\star, \precsim_i)$: if for a criterion i and a satisfaction value k, the value x is approved, then any value $x' \succsim_i x$ must also be approved.

- Clauses ϕ_α^{C2} ensure that approved sets are ordered by a set inclusion according to their satisfaction level: if an alternative x is approved at satisfaction level k according to criterion i, it should also be approved at satisfaction level $k' < k$.

- Clauses ϕ_α^{C3} ensure that each set of sufficient coalitions $\mathcal{T}$ is an upset for inclusion: if a coalition B is deemed sufficient at satisfaction level k, then a stronger coalition $B' \supset B$ should also be deemed sufficient at this level.

- Clauses ϕ_α^{C4} ensure that a set of sufficient coalitions are ordered by inclusion according to their satisfaction level: if a coalition B is deemed insufficient at satisfaction level k, it should also be at any level $k' > k$.

- Clauses ϕ_α^{C5} ensure that each alternative is not approved by a sufficient coalition of criteria at an satisfaction level above the one corresponding to its assigned category.

- Clauses ϕ_α^{C6} ensure that each alternative is approved by a sufficient coalition of criteria at a satisfaction level corresponding to its assignment.

Model variants.   As discussed in Section 2.3.2.3, the NCS model has many variants. Φ_α^C can easily be modified to account for two popular restrictions of the model, namely U^B-NCS (Unique profiles) and U^C-NCS (Unique set of sufficient coalitions), for more details see [Belahcene et al., 2018a; Tlili et al., 2022] .

3.3.1.2   A compact formulation-based on Pairwise Separation

A second formulation was introduced in [Belahcene et al., 2018b]. It leverages the fact that the partial inverse problem for NCS where *the approved sets are given* is much easier to solve and proposes a characterization of its feasibility based on pairs of alternatives. This approach leads to a compact formulation of the problem, with $\mathcal{O}(p \times |\mathcal{N}| \times |\mathbb{X}^\star|^2)$ variables and clauses. In addition, an extension of this formulation to the case of multiple categories was proposed in [Tlili et al., 2022].

To ease the readability, we expose in this section only the formulation in the case of two categories. For the case of multiple categories, we refer the reader to [Tlili et al., 2022].

In the following, we suppose given a set of reference alternatives $\mathbb{X}^\star$, an assignment $\alpha : \mathbb{X}^\star \to \{ \text{ GOOD }, \text{ BAD } \}$, and a tuple of accepted values $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{|\mathcal{N}|}$ such that, for each point of view $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$.

Observably sufficient and insufficient coalitions. Consider the sets of coalitions defined by

$$\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha) \quad := \quad cl^{\supseteq}_{\mathcal{P}(\mathcal{N})} \left(\bigcup_{g \in \alpha^{-1}(\text{ GOOD })} \{\{i \in \mathcal{N} : g \in \mathcal{A}_i\}\} \right), \quad (3.1)$$

$$\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) \quad := \quad cl^{\subseteq}_{\mathcal{P}(\mathcal{N})} \left(\bigcup_{b \in \alpha^{-1}(\text{ BAD })} \{\{i \in \mathcal{N} : b \in \mathcal{A}_i\}\} \right). \quad (3.2)$$

Any coalition in $\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is a superset of the set of criteria according to which some GOOD alternative is accepted and should, therefore, be accepted. Thus, $\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is a *lower bound* of the set of sufficient coalitions for any solution of Inv-NCS. Conversely, any coalition in $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is a subset of the set of criteria according to which some BAD alternative is accepted and should, therefore, be rejected. Thus, $\mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is an *upper bound* of the set of sufficient coalitions for any solution of Inv-NCS.

Characterization of solutions of Inv-NCS. The parameter $(\langle \mathcal{A}_i \rangle, \mathcal{T})$ is a solution of the instance α of Inv-NCS if and only if:

$$\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha) \subseteq \mathcal{T} \subseteq \mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) \quad (3.3)$$

Note that this equation allows characterizing the positive instances of Inv-NCS without referring to the set of sufficient coalitions of a solution, solely by checking if the sets $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ are disjoint. This leads to the following efficient characterization, based on the notion of *pairwise separation*.

Theorem 3.1. *An assignment α of alternatives to categories can be represented in the Non-Compensatory Sorting model if, and only if, there is a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{|\mathcal{N}|}$ such that:*

1. *(Upset) for each point of view $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$; and*

2. *(Pairwise separation) for each pair of alternatives $(g, b) \in \alpha^{-1}(\text{ GOOD }) \times \alpha^{-1}(\text{ BAD })$, there is at least one point of view $i \in \mathcal{N}$ such that $g \in \mathcal{A}_i$ and $b \notin \mathcal{A}_i$.*

This theorem provides a polynomial certificate for the positive instances of the Inv-NCS problem, thus proving its membership to the *NP* complexity class as a corollary.

The SAT formulation based on *pairwise separation* corresponds to the SAT encoding of both conditions of Theorem 3.1 [Belahcene et al., 2018b]. The first condition which ensures the monotonocity of scales is represented by a single family of clauses and operates on the same variables as the SAT formulation based on coalitions. In the second condition, additional binary variables are defined in order to represent the separation

between the alternatives. A unique family of logical clauses represents the separation concept of the theorem and additional clauses and binary variables are required in order to express this representation in SAT language.

Variables. Similarly to the formulation Φ_α^C described in the previous section, the formulation Φ_α^P operates on two types of variables.

- 'a' variables, representing the approved sets, with the exact same semantics as their counterpart in Φ_α^C,

- auxiliary 's' variables, indexed by a criterion $i \in \mathcal{N}$, an alternative g assigned to GOOD and an alternative b assigned to BAD , assessing if the alternative g is positively separated from b according to criterion i

Clauses. The formulation Φ_α^P is the conjunction of four types of clauses: ϕ_α^{P1} ensuring each $\mathcal{A}_i$ is an upset, ϕ_α^{P2} ensuring $[s_{i,g,b} = 1] \Rightarrow [g \in \mathcal{A}_i]$, ϕ_α^{P3} ensuring $[s_{i,g,b} = 1] \Rightarrow [b \notin \mathcal{A}_i]$, and ϕ_α^{P4} ensuring each pair (g,b) is positively separated according to at least one criterion.

It should be noted that, should ϕ_α^P be satisfiable, the set $\mathcal{T}$ of sufficient coalitions is not uniquely identified by the values of 'a' and 's' variables of one of its models. Indeed, if $\langle a_{i,x}\rangle, \langle s_{i,g,b}\rangle$ is an antecedent of 1 by ϕ_α^P, then the parameter $\omega = (\langle \mathcal{A}_i\rangle, \mathcal{T})$ with accepted sets defined by $\mathcal{A}_i = \{x \in \mathbb{X} : a_{i,x} = 1\}$ and any upset $\mathcal{T}$ of $(\mathcal{P}(\mathcal{N}), \subseteq)$ of sufficient coalitions containing the upset $\mathcal{S}_{\langle \mathcal{A}_i\rangle}(\alpha)$ and disjoint from the lower set $\mathcal{F}_{\langle \mathcal{A}_i\rangle}(\alpha)$ is a solution of this instance. Therefore, among the sets of sufficient coalitions compatible with the values of 'a' and 's' variables, we can identify two specific ones, $\mathcal{T}_{max}$ and $\mathcal{T}_{min}$.

Model variants. Φ_α^P can easily be modified to account for two popular restrictions of the model, namely U^B-NCS (Unique profiles) and U^C-NCS (Unique set of sufficient coalitions), in both cases two and multiple categories. For more details see [Tlili et al., 2022].

3.3.2   MaxSAT relaxations for Inv-NCS

The previous section introduced mathematical and computational tools addressing the *decision* problem: can a given assignment be represented in the Non-Compensatory Sorting model (or one of its variants)? However, such tools are not suited to the problem of learning a suitable NCS model from real data, because it does not tolerate the presence of noise in the data. There are several reasons for the input data not to reflect perfectly the model, e.g. imperfections in the assessment of performance

according to some point of view; mistaken assignment of an alternative to a category; or simply the oversimplification of reality presented by the model.

We addressed this issue by providing a relaxation of the decision formulations: instead of finding an NCS model restoring all examples of the learning set, we try to find the model that restores the most. We formulate the relaxed *optimization* problem of finding the subset of learning examples (reference alternatives together with their assignment) correctly restored of maximum cardinality with a *soft constraint* approach, using the language of weighted MaxSAT. This framework, derived from the SAT framework, is based on a conjunction of clauses $\bigwedge c_i$ where each clause c_i is given a non-negative weight w_i, and maximizes the total weight of the satisfied clauses.

To translate exactly our problem in this language, we leverage two basic techniques: we introduce switch variables 'z' allowing to precisely monitor the soft clauses we are ready to see violated, as opposed to hard clauses that remain mandatory; and we use big-stepped tuples of weights $w_1, \ldots, w_k$ with $w_1 \gg \cdots \gg w_k$ allowing to specify lexicographically ordered goals in an additive framework. The MaxSAT relaxation was proposed for both approaches: based on coalitions and based on pairwise separation conditions, and for each model variants (U^B-NCS and U^C-NCS) as well. We also generalize the formulation to the case of multiple categories. For more details, we refer the reader to [Tlili et al., 2022].

3.3.3 SAT/MaxSAT for Inv-NCS: main experimental insights

In addition to the work of formalizing learning algorithms, we were interested in the question of their efficiency. To account for this, several empirical studies were conducted. First, we conducted experiments to measure the performance regarding computation time by the size of the learning set. Second, we made a comparison with the stat of the art techniques. The experimentation protocol and the detailed results can be found in [Belahcene et al., 2018a]. Finally, we conducted other experiments to compare the different formulations [Tlili et al., 2022].

We enumerate eight of them, depicted in Figure 3.1 and specified by three binary parameters:

- the Non-Compensatory Sorting model of preference sought, either with a *unique boundary/limiting profile* (subscript $\mathbf{U^B}$), or with a *unique* set of sufficient *coalitions* (subscript $\mathbf{U^C}$) (see NCS variants in Sect. 2);

- the representation strategy adopted, based either on the explicit representation of the *coalitions* of criteria (superscript $\mathbf{C}$) or on the *pairwise* separation of alternatives (superscript $\mathbf{P}$); and

- the problem description, either *deciding* whether an instance can be represented
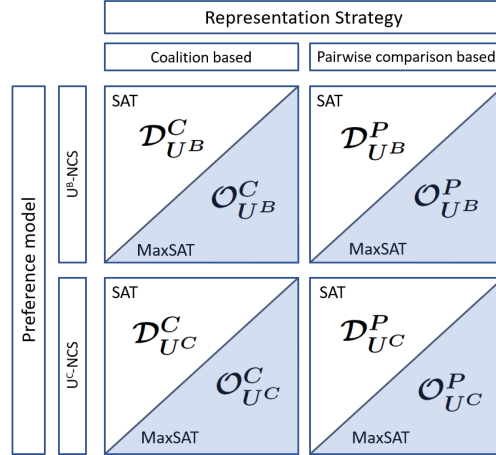
Figure 3.1: Approaches for comparing learning algorithms

in the model ($\mathcal{D}$) with a SAT solver, or *optimizing* the ability of the model to represent the assignment ($\mathcal{O}$) with a MaxSAT solver.

The details of the experimental protocol and results' discussions can be found in [Tlili et al., 2022]. From these experiments, we were able to conclude that the separation-based representation proposed for learning U^B and U^C models is at least as good as the coalition-based one in terms of generalization and for both types of preference information (perfect and not-so-perfect preferences). The computation time of the two representations evolves depending on the number of reference alternatives and the number of criteria; the separation-based representation performs better when the number of criteria increases, while it is not the case when the number of reference alternatives increases. Increasing the number of categories penalizes the separation-based representation proposed for learning the U^B model since the number of clauses depends quadratically on the number of categories.

However, for real-world decision problems, assuming that the number of reference assignments is ~100 examples, we can consider two types of applications: an application that involves a large number of criteria ($|\mathcal{N}| >\sim 12$) and therefore the separation-based representation seems better as it is faster and generalizes better than the first one, and an application that involves a limited number of criteria ($|\mathcal{N}| <\sim 10$), in this case, the coalition-based representation is slightly faster and generalizes less than the separation-based one. Finally, our work shows that, when learning MCDA models from preference information, SAT and MaxSAT languages can be relevant and efficient. This is precisely the case for ordinal MCDA aggregation procedures based on a pairwise comparison of alternatives (so-called outranking methods, see [Figueira et al., 2005]).

3.4 Learning NCS Model Parameters: new perspectives

In the previous section, we presented devices for eliciting the parameters of sorting models indirectly from a set of assignment examples, i.e., a set of alternatives with corresponding desired categories. To be applied, such preference learning approaches make some assumptions about the structure of the criteria.

On the one hand, in MCDA, preference elicitation methods require a *preference order* on each criterion. Such preference order results from the fact that alternative evaluations/scores correspond to maximized performances (profit criterion) or minimized (cost criterion), resulting in monotone preference data. In multicriteria sorting problems, this boils down to a higher evaluation on a profit criterion (on a cost criterion, respectively) favors an assignment to a higher category (to a lower category, respectively). However, there are numerous situations where the criteria evaluation is not related to category assignment in a monotone way. For instance, consider Example 3.1 for illustration.

Example 3.1.

A computer-products retail company is distributing a new Windows tablet, and wants to send targeted marketing emails to clients who might be interested in this new product. To do so, clients are to be classified into two categories: *potential buyer* and *not interested*. To avoid spamming, only clients in the former category will receive an email. To sort clients, four characteristics are considered as criteria, all of them being homogeneous to a currency e.g. € : the turnover over the last year of *(i)* Windows PC, *(ii)* Pack Office, *(iii)* Linux PC, and *(iv)* Dual boot PC.

The aim of the company is to advertise a new Windows tablet. Thus, both first two criteria are to be maximized (the more a client buys Windows PCs and Pack Office, the more he is interested in products with a Windows system), and the third criterion is to be minimized (the more a client buys Linux PCs, the less he is interested in products with a Windows system). The marketing manager is convinced that the last criterion should be taken into account, but does not know whether it should be maximized or minimized; a subset of clients has been partitioned into not interested/potential buyer.

Considering situations like the one described by Example 3.1, the goal of the learning task is to simultaneously learn the classifier parameters and the preference direction (profit or cost) for the last criterion. More generally, the idea is to consider that the preference order on each criterion is *unknown*, i.e. the evaluations of alternatives

induce monotone preferences, but the preference directions on criterion are unknown (i.e. whether each criterion is maximized or minimized).

The second assumption refers to the fact that the preferences on criteria are *not necessarily monotone* but possibly *single-peacked (or single-valley)*. For instance, consider Example 3.2 for illustration.

Example 3.2.

Consider a veterinary problem in cattle production. A new cattle disease should be diagnosed based on symptoms: each cattle should be classified as having or not having the disease. New scientific evidence has indicated that substance A in the animal's blood can be predictive in addition to usual symptoms. Still, there is no clue how the level of substance A should be considered. Does a high, a low level, or a level between bounds of substance A indicate sick cattle?

The veterinarians' union has gathered many cases and wants to benefit from this data to define a sorting model based on usual symptom criteria and the level of substance A in the animal's blood. Hence, the sorting model should be inferred from data, even if the way to account for the substance A level is unknown.

In the previous example, it is unclear to the decision-maker how to account for the level of substance A in blood in the classification of alternatives (cattle, client). This example corresponds to a single-peaked criterion, i.e. criterion for which preferences are defined according to a "*peak*" corresponding to the best possible value; on such a criterion, the preference decreases with the distance to this peak. In other words, the peak corresponds to a target value below which the criterion is to be maximized, and above which the criterion is to be minimized. Such criteria are frequent in the medical domain (getting close to a normal blood sugar level) and chemical applications (get close to a neutral PH), ... It is also natural to consider the reverse side of the single-peaked preference that, is the *single-valley* preference (illustrated by a "V" curve). In such a case, the bottom is the less preferred value, and the more the values are far from the bottom, the more preferred they are.

Therefore, in our works, we focus on the MR-Sort model. Our concerns were twofold: (i) we simultaneously aim to uncover from a learning set the criteria preference directions and the MR-Sort parameters (criteria weights, limit profiles, majority threshold). Our proposals to answer this objective are summarized in Section 3.4.1; (ii) dealing with single-peaked and single-valley preferences no longer fit the scope of monotone preferences. Therefore, we intend to consider a more extensive scope, i.e. non-monotone preferences, since we want to learn MR-Sort models from possibly single-peaked/ single-valley preferences. The proposals to account for this are summarized in Section 3.4.2.

3.4.1 Learning MR-Sort models with latent criteria ciretion

To account for the learning of the preference direction in the Inv-MR-Sort problem, we based our proposal on the heuristic proposed by [Sobrie, 2016; Sobrie et al., 2019]. The heuristic is an evolutionary population-based algorithm and learns an MR-Sort model that best matches a learning set composed of assignment examples. Each individual in the population is an MR-Sort model, i.e., values for limit profiles b^h, criteria weights w_i, and the majority level λ; each individual is denoted by $(\langle b \rangle, w, \lambda)$. After an initialization step that generates the first population, the algorithm proceeds to evolve the population of MR-Sort models iteratively until a model in the population perfectly restores the learning set or a maximum number of iterations is reached. Moreover, at each iteration, the algorithm tries to improve the fitness of each MR-Sort model in the population (the proportion of correctly restored examples in the learning set) by performing two consecutive steps: (i) optimize the weights and majority level (limit profiles being fixed) using linear programming (LP), and (ii) improve heuristically the limit profiles (weights and majority level being set). The 50% best models are kept in the population for the next iteration, while 50% new MR-Sort models are randomly generated.

The works of [Sobrie, 2016; Sobrie et al., 2019] assume the monotonicity of criteria in the MR-Sort model to be learned. More precisely, the definition of the Inv-MR-Sort problem assumes, without loss of generality, that the decision-maker preferences are increasing with the criteria performances (the greater, the better). Therefore, within the thesis of Minoungou [2022], we investigated the possibility of extending the Inv-MR-Sort problem to the case where preferences are still monotone, but the criteria preference directions are not known, i.e., we do not know whether the criteria are to be maximized or minimized. We implemented two approaches:

- The first one, titled *duplication-based*, relies on the heuristic of [Sobrie, 2016] at two consecutive phases. The first one is for learning the preference directions, and the second takes the learned directions as input and mobilizes the heuristic again for learning the other parameters of the model (profiles, weights and majority threshold) [Minoungou et al., 2020].

- The second approach, titled *mixed-based*, extends the heuristic to learn the preference direction simultaneously with the other MR-sort parameters. It consists of evolving models with both gain and cost criteria in the population of models during the learning process.

Although each has advantages and shortcomings, the experiments have demonstrated that the first method is the most effective. Therefore, we choose to briefly describe it in what follows.

3.4.1.1   Duplication-based approach

The first approach to determine the criteria preference directions combines two consecutive steps. Each step is based on the heuristic of [Sobrie, 2016], with additional adjustments. The idea is to start by resolving an MR-Sort problem by duplicating the subset of criteria Q ($Q \subseteq \mathcal{N}$ and $|Q| = q$) whose preference direction is unknown into an identical Q' set, such that the criteria in Q have an increasing preference direction. Those in Q' a decreasing one. The intuition behind the duplication is to foster the algorithm to inhibit the criterion with the "incorrect" preference direction while making the other criterion influential. Therefore, the main steps of the methodology are as follows:

1. **Learning the q preference directions**. It consists in resolving an Inv-MR-Sort problem with $n+q$ criteria, such that n is the initial number of criteria and q is the number of criteria whose preference direction is unknown. Solving this problem with the heuristic will allow us to learn the parameters: b (of dimension $n + q$), w (of $n + q$ criteria) and the threshold λ.

2. **Retrieving the preference direction of the q latent criteria**. The idea is given a couple (i, j) of criteria ($i \in Q$, $j \in Q'$ and j is the duplication of i); we analyze each criterion's weight to retrieve the right direction. Three situations are considered: (i) both weights are equal to zero, (ii) both are different to zero, and (iii) one of them is zero, and the other is not. For instance, in the last situation ($w_i = 0$ or $w_j \neq 0$, or vice versa), we keep the direction of the criterion whose weight is not zero. Situation (ii) is the most tricky one. To fix the preference direction, we ground our analysis on the position of profiles b regarding the endpoint of the scales $\mathbb{X}_i$ and $\mathbb{X}_j$. The intuition is that profiles on criterion i (or j) close to the endpoints of the scale X_i (or $\mathbb{X}_j$) indicates that criterion i (or j) is "inhibited". Therefore, we select the preference direction corresponding to criterion i or j as the one for which the profile is further away from the endpoints of the scales $\mathbb{X}_i$ and $\mathbb{X}_j$ (we refer the reader to [Minoungou et al., 2020] for more details).

3. **Learning the standard MR-sort parameters**. Once the q preference direction criteria are fixed from the last step, it consists in resolving a classical Inv-MR-Sort problem with n criteria. For this, we reduce the problem with $n + q$ criteria to a problem with n criteria and resolve this latter with the heuristic in [Sobrie et al., 2019] to learn the final parameters' values of the MR-Sort problem.

3.4.1.2   Main experimental insights

To analyze the behavior of the approach, we conducted several experimental analyses to measure: i) Regarding the computing time, how the algorithm copes with large

datasets, ii) the ability of the algorithm to restore a dataset when criteria preference direction are latent, iii) how many assignment examples should the learning set contains so that learned model accurately classify new alternatives, iv) How does the algorithm cope with noisy datasets (i.e. alternatives falsely assigned to wrong categories).

The extensive numerical simulations demonstrate the capability of the algorithm to correctly estimate both preference direction and the other model parameters with an accuracy of over 90% (for a noise-free learning set of 250 examples). Moreover, the algorithm showed to be robust in the case of noisy data. Finally, the proposed solution features a very contained computational complexity both in the training and inference phases.

3.4.2 Learning MR-Sort models with single-peaked preferences

Another situation in which the current preference learning tools within the MCDA context are not satisfactory is when the preferences on criteria are not necessarily monotone. We seek to provide efficient means to solve the Inv-MR-Sort problem with single-peaked preference criteria.

Indeed, the standard approach in the MCDA literature is to carefully craft the set of evaluation criteria so that these criteria are to be either maximized (gain criterion) or minimized (cost criterion). This boils down to the hypothesis that the data have a monotonic property. Our approach is relaxing this hypothesis allowing the criteria to be cost, gain, single-peaked or single-valley criteria. Some works account for the non-monotonicity of preferences in value-based models (see, e.g. [Despotis and Zopounidis, 1995]). Our work aimed to extend this idea of non-monotone criteria to outranking methods and, in particular, to the MR-Sort model (see Chapter 2). Specifically, we tackled the problem of inferring, from a dataset (learning set), an MR-Sort with possibly non-monotone criteria. The challenge is that this inference problem is already known to be difficult with monotone criteria, see [Leroy et al., 2011].

Before exposing our contributions, we first describe in what follows how we can formalize non-monotone criteria in an MCDA context. More precisely, we considered single-peaked and single-valley criteria.

Let us denote $\mathbb{X}_i$ the finite set of possible values on criterion i, $i \in \mathcal{N} = \{1, \ldots, n\}$; we suppose w.l.o.g. that $\mathbb{X}_i = [min_i, max_i] \subset \mathbb{R}$. In an MCDA perspective, single-peaked criteria (and single-valley criteria) can be interpreted as "locally-monotone" criteria, as they are to be maximized (a cost criterion to be minimized, respectively) below the peak p_i, and as a cost criterion to be minimized (a gain criterion to be maximized, respectively) above the peak p_i (see Def 3.1). We choose to model single-peaked (single-valley) preferences, as they remain locally monotone and therefore "close"

to the structured perspective of MCDA. Note also that single-peaked and single-valley preferences embrace the case of gain and cost criteria: a gain criterion corresponds to single-peaked preferences when $p_i = max_i$ or single-valley preferences with $p_i = min_i$, and a cost criterion corresponds to single-peaked preferences when $p_i = min_i$ or single-valley preferences with $p_i = max_i$.

Definition 3.1. *Preferences $\succcurlyeq_i$ on criterion i are:*

- *single-peaked preferences with respect to $\geq$ iff there exists $p_i \in \mathbb{X}_i$ such that: $x_i \leq y_i \leq p_i \Rightarrow p_i \succcurlyeq_i y_i \succcurlyeq_i x_i$, and $p_i \leq x_i \leq y_i \Rightarrow p_i \succcurlyeq_i x_i \succcurlyeq_i y_i$*

- *single-valley preferences with respect to $\geq$ iff there exists $p_i \in \mathbb{X}_i$ such that: $x_i \leq y_i \leq p_i \Rightarrow p_i \succcurlyeq_i x_i \succcurlyeq_i y_i$, and $p_i \leq x_i \leq y_i \Rightarrow p_i \succcurlyeq_i y_i \succcurlyeq_i x_i$*

If we go back to our question, which is about learning MR-Sort parameters with single-peaked preferences, the first step is to be able to represent a single-peaked preference. Indeed, from the previous definition, one can see that the approved sets ($\mathcal{A}_i$) can not be represented using frontiers between consecutive categories. However, approved sets should be compatible with preferences, i.e. such that:

$$\begin{cases} x_i \in \mathcal{A}_i^h \text{ and } x_i' \succcurlyeq_i x_i \Rightarrow x_i' \in \mathcal{A}_i^h \\ x_i \notin \mathcal{A}_i^h \text{ and } x_i \succcurlyeq_i x_i' \Rightarrow x_i' \notin \mathcal{A}_i^h \end{cases} \tag{3.4}$$

In case of a single-peaked criterion with peak p_i, we have:

$$\begin{cases} x_i \in \mathcal{A}_i^h \text{ and } p_i \leq x_i' \leq x_i \Rightarrow x_i' \in \mathcal{A}_i^h \\ x_i \in \mathcal{A}_i^h \text{ and } x_i \leq x_i' \leq p_i \Rightarrow x_i' \in \mathcal{A}_i^h \\ x_i \notin \mathcal{A}_i^h \text{ and } p_i \leq x_i \leq x_i' \Rightarrow x_i' \notin \mathcal{A}_i^h \\ x_i \notin \mathcal{A}_i^h \text{ and } x_i' \leq x_i \leq p_i \Rightarrow x_i' \notin \mathcal{A}_i^h \end{cases} \tag{3.5}$$

Therefore it appears that with a single-peaked criterion with peak p_i, the approved sets $\mathcal{A}_i^h$ can be specified by two thresholds $\overline{b}_i^h, \underline{b}_i^h \in X_i$ with $\underline{b}_i^h < p_i < \overline{b}_i^h$ defining an interval of approved values: $\mathcal{A}_i^h = [\underline{b}_i^h, \overline{b}_i^h]$. Analogously, for a single-valley criterion with peak p_i, the approved sets $\mathcal{A}_i^h$ can be specified using $\overline{b}_i^h, \underline{b}_i^h \in X_i$ (such that $\underline{b}_i^h < p_i < \overline{b}_i^h$) as $\mathcal{A}_i^h = X_i \setminus]\underline{b}_i^h, \overline{b}_i^h[$

Given a single-peaked criterion i for which the approved set is defined by the interval $\mathcal{A}_i^h = [\underline{b}_i^h, \overline{b}_i^h]$, consider the function $\phi_i : X_i \to X_i$ defined by $\phi_i(x_i) = |x_i - \frac{\overline{b}_i^h + \underline{b}_i^h}{2}|$, i.e., the absolute value of $x_i - \frac{\overline{b}_i^h + \underline{b}_i^h}{2}$. Then, the approved set can be conveniently rewritten as : $\mathcal{A}_i^h = \{x_i \in X_i : \phi_i(x_i) \leq \frac{\overline{b}_i^h - \underline{b}_i^h}{2}\}$. In other words, when defining approved sets, a single-peaked criterion can be re-encoded into a cost criterion, evaluating alternatives as the distance to the middle of the interval $[\underline{b}_i^h, \overline{b}_i^h]$, and a frontier corresponding to

half the width of this interval. Analogously, the same reasoning can be applied to a single-valley criterion.

With this definition of approved sets, we proposed two approaches for learning the MR-sort models with single-peaked criteria, described in the following.

An exact approach. We aim to learn the parameters of an MR-Sort model with potentially single-peaked criteria from assignment examples. Our learning process consists of the resolution of a Mathematical Integer Program (MIP) based on L, the set of assignment examples (the learning set). For recall it corresponds to the couple $(A^*, \mathcal{C})$, where $\mathcal{C} = \{cat(a), \forall a \in A^*\}$; that is each alternative $a \in A^* \subset \mathbb{X}$ is assigned to a desired category $cat(a) \in \{1, \ldots, p\}$. Therefore we call the new Inverse MR-Sort problem *Inv-MR-Sort-SP* problem since we consider single-peaked/single-valley criteria.

In this problem, we assume *not knowing* in advance the type of preferences of criteria involved in the learning process. In addition, as said previously, we consider single-peaked and single-valley criteria. Moreover, we treat the case with two categories. Thus, we denote by $\mathcal{S}$ the set of single-peaked and single-valley criteria, and $s, s = |\mathcal{S}| \leq n$ the number of single peaked and single-valley criteria. We also denote by $\mathcal{Q}$ the set of criteria with unknown preference directions, and $q, q = |\mathcal{Q}| \leq n$ the cardinal of this set. We note $IMSS_{q|n}$ the *Inv-MR-Sort-SP* problem with q, the number of criteria with unknown preferences directions, and n the number of criteria which possibly contains some single-peaked/single-valley criteria.

The resolution process will take as input a learning set containing assignment examples and computes:

- the nature of each criterion (either cost, gain, single-peaked, or single-valley criterion),

- the weight w_i attached to each criterion $i \in \mathcal{N}$, and an associated majority level λ,

- the frontier between category C^h and C^{h+1}, i.e. the value b_i^h if criterion i is a cost or a gain criterion, and the interval $[\underline{b}_i^h, \overline{b}_i^h]$ if criterion i is a single-peaked or single-valley criterion.

The technical details of the MIP are described in [Minoungou et al., 2022]. Finally, experiments on randomly generated instances give us the following insights. Although exact methods are typically computationally intensive, the computation time is relatively affordable for medium-sized models (less than 3 minutes for 200 alternatives in the learning set and up to $n = 9$ and $q = 4$ in the model when the timeout is set to 1 hour). The computation time could be reduced as our experiments were performed with a limited number of threads set to 10. Moreover, the algorithm can restore

accurately new assignment examples based on the learned models (0.93 on average up to 9 criteria) and remains relatively efficient regarding the number of criteria with unknown preference directions.   Finally, the restoration rate of criteria preference direction correlates with such criteria importance in the model.   The preference directions of criteria with importance below $\frac{1}{2n}$ are the most difficult to restore. These results are valid with a fixed-size learning set (200).

Our experiments give good results, except they are limited by the model's size, which becomes rapidly intractable (200 alternatives, four criteria). Experiments suggest that the correct restoration of criteria preference directions requires datasets of significant size. To account for this, we follow a heuristic-based approach which is tractable with large datasets. See the following point.

A heuristic-based approach.   To cope with the tractability problem of the exact approach, a heuristic approach is proposed, which is an adaptation of the evolutionary metaheuristic of [Sobrie et al., 2013] (sorting into two categories). The tricky point, which requires adaptation, is to evolve not the level of the limit profile but the two extremities of the interval of approved values. In other terms, we assume that the directions of the criteria (monotonic or non-monotonic) are known in advance, and the "acceptable" values of the categories are in the form of intervals. The goal is to learn the values of the profile intervals $[\underline{b}_i, \bar{b}_i]$).

Two versions are proposed. The first consists of randomly and successively learning the first and then the second interval value of the profiles of single-peaked criteria. The second variant consists in learning both interval values of single-peaked criteria simultaneously. We refer the reader interested to [Minoungou, 2022] for the technical details.

The result of the experiments (on artificial instances) is that the two variants lead to approximately equal classification qualities. The second variant leads to computation times that increase strongly with the size of the learning set. The rest of the experiment is therefore carried out with the first variant. The results are convincing both on free noise data and noisy data. The algorithm is also applied to ASA[1] data [Lazouni et al., 2013], where the range of values approved for the "glycaemia" criterion seems to be well detected. Two real datasets, from the UCI Repository [2] [Cortez et al., 2009], relating to the assessment of wines by experts are also dealt with; the wines being described by some of their chemical characteristics. The classification quality of the algorithm is comparable to that obtained with a Support Vector Machine (SVM) technique (for expert assessments partitioned into two categories in three different ways). This result seems encouraging for the rest of our work on non-monotone data.

---

[1] ASA stands for "*American Society of Anesthesiologists*".
[2] https://archive.ics.uci.edu/ml/datasets/wine

3.5   Summary

Preference handling and elicitation are crucial in many computer science domains, including recommender systems, interface customization and personal assistants [Peintner et al., 2008]. Our research works in this line seek to advance state-of-the-art with new tools borrowed from AI (Boolean-based formulations) and tackle new problems, such as learning with non-monotone preferences.

Finally, in addition to NCS and its variants, we have considered other models and decision problems. Typically, we were interested in a method based on outranking relations, called Ranking based on Multiple reference Profiles (RMP) [Rolland, 2013]. The RMP model for ranking alternatives by the strength with which they outrank some underlying reference points or profiles has been introduced in Rolland [2008, 2013]. It has been axiomatically characterized in [Bouyssou and Marchant, 2013]. Real-world applications can be found in [Ferretti et al., 2018] or [Khannoussi et al., 2019].

More precisely, we contribute by proposing indirect elicitation procedures for the S-RMP method (where the importance relation on criteria coalitions is determined by additive weights), such that a decision-maker provides pairwise comparisons of alternatives from which the S-RMP preference parameters (weights, reference points, and the lexicographic order on reference points) are inferred.

We have proposed three different approaches. First, in [Liu, 2016] we formulate the elicitation of an S-RMP model as a Mixed Integer linear optimization problem (MIP). In this optimization program, the variables are the parameters of the S-RMP method and additional technical variables, which enable to formulate of the objective function and the constraints in a linear form. The aim is to minimize the Kemeny distance (see [Kemeny, 1959]) between the partial Ranking provided by the decision-maker (i.e. the comparisons) and the S-RMP ranking. The resolution of this optimization program guarantees that the elicited S-RMP model best matches the pairwise comparisons in terms of the Kemeny distance between the comparisons provided by the DM and the S-RMP ranking.

Second, a meta-heuristic was proposed to indirectly elicit an S-RMP model from pairwise comparisons in [Liu et al., 2014; Liu, 2016]. Unlike the MIP version, this metaheuristic does not guarantee that the inferred model is the one which minimizes the Kemeny distance to DM's statements. Indeed, the perspective is obtaining an S-RMP model that fits the decision maker's comparisons "well" within a "reasonable" computing time. This metaheuristic is based on an evolutionary algorithm in which a population of S-RMP models is iteratively evolved.

The algorithms mentioned above suffer, however, from limitations:

- both algorithms only consider an additive representation of the criteria importance relation, which can be restrictive when the interaction between criteria occurs;

- the MIP-based approach implies computational difficulties in dealing with datasets whose size corresponds to real-world decision problems

- the heuristic approach is fast but cannot always restore an S-RMP model compatible with a set of comparisons whenever it exists.

To circumvent these limitations, we proposed to rely on SAT/MaxSAT formulations which are computationally efficient to tackle the learning task of the parameters of an RMP model. Our experimentation has addressed a real case study, showing that the approach is feasible also when applied to real data sets. This work is not described in this manuscript. For more detail, we refer the reader to [Belahcene et al., 2018c, 2022]

Now, our ambition is to continue to advance this line of research by deepening certain questions, exploring new decision models or even looking for new devices by taking advantage, for example, of the benefits of machine learning techniques in terms of efficiency and capacity to process large Dataset. See Chapter 5 for a discussion.

Supporting Decisions: a Panel of Explainability Tools

In the previous chapter, we addressed and summarized our contributions regarding providing efficient tools to learn preference models from the learning set to represent the decision-maker judgment faithfully. Establishing such a model will allow deriving recommendations to answer the decision-maker's problem. To enhance the trust of the DM towards these recommendations, we investigated the question of how and what supporting evidence to provide to justify such recommendations. One of the difficulties of this question is that the relevant concept of an explanation may differ depending on several aspects (for instance, the target audience, the form of the explanations). This chapter is devoted to summarizing our contributions to this topic.

4.1 Explainable Artificial Intelligence: Positioning

In recent years we have witnessed the emergence of new questions and concerns regarding AI-based systems. A new field under the name of "eXplainable AI (XAI)" has emerged [Gunning, 2017], with the mission of enlightening end-users on the functioning of these systems and providing answers to the "why" question. More precisely, the DARPA, at the origin of this buzz word, gives the following definition:

> "provide *users* with *explanations* that enable them to *understand* the system's overall *strengths and weaknesses*, convey an understanding of how it will behave in future or different situations, and perhaps permit users to *correct* the system's mistakes".

Moreover, the increasing need for AI explainability has also prompted governments to introduce new regulations. The most famous one is the General Data Protection Regulation (GDPR), which was introduced by the European Union in 2016 and has been enforced since 2018[1]. Since then different works were dedicated to analyzing this requirement from a legal point of view [Goodman and Flaxman, 2017; Wachter et al., 2017]. Finally, even if we are witnessing an explosion of work bearing interest in this question of explainability, notably in the field of Machine Learning (see, for

---

[1] European Council (2016). The general data protection regulation.

example, [Biran and Cotton, 2017; Guidotti et al., 2019; Mohseni et al., 2018; Barredo Arrieta et al., 2020], to cite a few), this question is not entirely new and goes back to expert systems [Swartout, 1983; Gregor and Benbasat, 1999], and since then many works have emerged. These works investigate a variety of issues, such as: generating and providing explanation [Carenini and Moore, 2006; Nunes et al., 2014]; identifying what the desirable features of an explanation are from the point of view of its recipient [Herlocker et al., 2000; Tintarev, 2007; Mohseni et al., 2021]. More recently, Miller [2019] discussed such issues from the point of view of philosophy, psychology, and cognitive science.

Finally, the concept of explanation in Artificial Intelligence (AI) may be described according to several key characteristics, including the *target audience*: end-user, domain expert, knowledge engineer, etc. [Barredo Arrieta et al., 2020; Mohseni et al., 2021], *the scope*: local vs global [Wick and Thompson, 1992; Doshi-Velez and Kim, 2017; Liao et al., 2020; Mohseni et al., 2021], *the type*: contrastive, counterfactual, etc. [Lipton, 1990; Miller, 2019; Gupta et al., 2022; Chandrasekaran et al., 1989], *the trigger*: action on a graphical interface, asking predefined textual questions,... [Swartout and Smoliar, 1987; Cashmore et al., 2019] and *the form* of the explanations: visual (images, graphs, etc.), verbal (template texts, naturally generated texts, etc.) [Simonyan et al., 2014; Mohseni et al., 2021; Poli et al., 2021]. It is not our ambition to make state of the art or discuss XAI's different works, definitions, or contributions. We refer the reader interested to the extensive literature on the subject. Our message is that the concept of explanation cannot be unique, and we cannot claim to have a generic explanation common to all applications and users.

Our work is part of the ambition of building systems accountable for their decisions. In decision-aiding, the task is difficult because this accountability demand may require the system to explain an internal reasoning process built during the interaction with the user. In particular, the system may have inferred some preferences of the user before using a specific model, which is considered adequate. As a result, such an explanation is prone to be challenged and even contradicted, leading to the revision of the recommendation rather than a failure of the process (see Chapter 5 for a discussion on the issues related to revision and challenging an explanation). We investigated the question of explainability within different domains: Multiple Criteria Decision Aiding [Belahcene, 2018; Amoussou, (in progress); Ouerdane, 2009], Rule-based systems [El Mernissi, 2017; Baaj, 2022; Baaj et al., 2021] and more recently optimization systems [Lerouge, (in progress)]. As we have chosen to focus this document on contributions related to MCDA, we will not detail in this chapter our contributions within the two other domains (see Chapter 5 for a brief discussion on our ongoing work on explainability for optimization systems).

Explainability in MCDA. In this context, our main concern is developing principle-based approaches and cognitively bounded models of explanations for *end-users*. By principle-based approach, we mean that each explanation is attached to a number of well-understood properties of the underlying decision model. By cognitively bounded, we mean that the statements composing an explanation will be constrained to remain easy to grasp by the receiver (decision-maker). More generally, we seek to answer the following question:

> "Given a decision model and a set of preference information, is there a principled way to define a simple complete explanation for a decision?"

To answer the previous question, in our various works, we essentially consider the following ingredients:

- The decision problem. We have devoted our work to studying and constructing explanation patterns for different decision problems: choice, pairwise comparison and assignment (see Chapter 2). Indeed, as the requirements vary significantly from situation to situation and from decision-maker to another, we do not believe in providing a unique type of explanation. Under such a perspective, we considered different decision models: weighted sum, additive utility, and the Non-Compensatory Sorting model (see Chapter 2).

- The collected (expressed) Preference Information (PI). Preference information, as we have seen in Chapter 2, is the essence of the decision problem. It represents the information provided by the decision-maker and is, therefore, an essential element both in the specification of the aggregation model and in the construction of the explanation.

- The explanation language. We aim to provide a formal language and reasoning machinery to support (explain) the output of a decision model. We build on the notion of *argument schemes*, that are stereotypical patterns of reasoning, which are used as presumptive justification for generating arguments. Each scheme is associated with a set of critical questions, which allow one to identify potential attacks on an argument generated by the scheme [Walton, 1996; Atkinson and Bench-Capon, 2021].

  In other terms, we can see a scheme as an operator tying a sequence of statements, called the premise, satisfying some conditions, into another statement called the conclusion. As we deal with preferences, argument schemes derive new preferences from previously established ones. As we shall see, in most of our proposals, an explanation takes the form of a pair ⟨premises, conclusion⟩, such that the premises are "minimal" and support the explanation.

- The approaches or techniques to compute explanations.   To identify such patterns, and depending on the situations, we have used different approaches and techniques, from mathematical programming to logic-based tools (SAT/MaxSat formulation, MUS).

Finally, in the different works we have carried out towards the formalization of the concept of explanation, we have considered various aspects in producing explanations when possible. More precisely, we were interested in:

- **Computation**: *How difficult is it to produce an explanation?* We expect this question to require notions and tools from the field of Computational Complexity.

- **Simplicity**: Although they are of a formal nature, the explanations produced should eventually be presented to humans. Thus, *Can we keep the explanations simple enough?* Neither natural language generation nor in vivo experimentation belong to the scope of our contributions, so the complexity of explanations shall be assessed through proxies, such as the length or number of elements that make up the explanation.

- **Completeness**: *Can we explain every 'true' result, that can be deducted from the preference information and the model?*

- **Soundness**: *Could we explain 'false' results, claiming the impossibility of an event that could happen or the possibility of an event that cannot happen?*

4.2   Explaining Recommendations Stemming from MCDA Models

While elicitation describes operations that formalize the knowledge of preferences, explanations focus on establishing a relation between the obtained preference model and the user (decision-maker). This chapter tells the story of our different works on explainability in the context of multiple criteria decision aiding. The work presented here results from long collaborations with several colleagues and PhD students [Belahcene, 2018; Amoussou, (in progress)]. Collaborations that go back to my PhD thesis [Ouerdane, 2009][2]. The results of these different collaborations for different decision problems and models are summarized in Table 4.1.

In the rest of this chapter, we have chosen to present the various contributions through examples and limit the technical details to ease the understanding. Readers interested in the technical details are invited to consult published articles attached to each contribution (see Appendices C).

---

[2]That's to say that it's been a long time...!

Decision Problem	Model	Reference
Choice	Weighted Majority	[Labreuche et al., 2011]
	Additive Utility	[Labreuche et al., 2012]
Pairwise Comparison	Additive Utility	[Belahcene et al., 2019] [Belahcene et al., 2017a]
Sorting	NCS	[Belahcene et al., 2018b], [Belahcene et al., 2017b]

Table 4.1: Our contributions for explainable MCDA

4.2.1 Explaining a recommended choice

Our first contributions for explaining recommendations stemming from MCDA model concern explaining a recommended choice. These works result from collaborations with Christophe Labreuche (Thales Research and Technology) and Nicolas Maudet (LIP6, Sorbonne Université).

The decision model we rely on is based on the *Weighted Condorcet principle*: options are compared in a pairwise fashion, and an option a is preferred to an option b when the cumulated support that a is better than b outweighs the opposite conclusion. We proposed two different approaches for explaining a recommended choice with different assumptions: (i) a single value for the weight vector (see Section 4.2.1.1), and (ii) a set of vectors compatible with the PI (see Section 4.2.1.2).

4.2.1.1 Explanation when PI is complete

In this work, we seek to provide simple but complete explanations for the fact that a given option is a Weighted Condorcet Winner (WCW)[3], by considering two types of PI: (i) the importance of the criteria, and (ii) the ranking of the different options (linear orders). To illustrate the problem, let us consider the following situation:

Example 4.1. [Labreuche et al., 2011]

There are 6 options $\{a, b, c, d, e, f\}$ and 5 criteria $\{1, \cdots, 5\}$ with respective weights as indicated in the following table. The (full) orderings of options must be read from the top (first rank) to the bottom (last rank).

---

[3]Of course, a strong assumption here is that a WCW exists. This assumption is removed in the next section.

criteria	1	2	3	4	5
weights	0.32	0.22	0.20	0.13	0.13
ranking	c	b	f	d	e
	a	a	e	f	b
	e	f	a	b	d
	d	e	c	a	f
	b	d	d	c	a
	f	c	b	e	c

In the previous situation, option a is the WCW, but it does not come out as an obvious winner, hence the need for an explanation. Of course, a possible explanation is always to explicitly exhibit the computations of every comparison, but even for a moderate number of options, this may be tedious. Thus, a tentative "natural" explanation that a is the WCW would be as follows:

Example 4.2. (Ex. 4.1 Cont.)

- First consider criteria 1 and 2, a is ranked higher than e, d, and f in both, so is certainly better.

- Then, a is preferred over b on criteria 1 and 3 (which is almost as important as criterion 2).

- Finally, it is true that c is better than a on the most important criterion, but a is better than c on all the other criteria, which together are more important

---

Of course, our aim was not to produce such natural language explanations but to provide the theoretical background upon which such explanations can later be generated. Thus, to construct such an explanation, we have considered different ingredients regarding both the expression of the preferences among options and the weights of criteria. These ingredients correspond to the *elementary chunks* that we allow being used in the formulation of the explanation to meet the need for intelligible, relevant and cognitively simple explanations. On the one hand, we need statements to express preferences: a set of *basic preference statements* (a preference between two options regarding a given criterion), a set of *factored preference statements* (preference of an option over a subset of options on a given criterion, or preference of an option over a subset of options on a subset of criteria), and a set of *importance statements* (to specify the weight of a criterion). Moreover, we may have different types for each preference statement: negative (against the WCW), positive (in favor of the WCW) and neutral. These different types

are illustrated in Example 4.3.

> **Example 4.3. (Ex. 4.1 Cont.)**
>
> Basic preference statements: $[1 : c \succ a]$ (negative), $[1 : c \succ f]$ (neutral), $[1 : a \succ e]$ (positive).
>
> Factored preference statements: $[1 : c \succ a, e]$ (negative), $[1, 2 : e \succ d]$ (neutral), and $[1, 2 : a \succ d, e, f]$ (positive).

On the other hand, we seek for a complete and minimal explanation. By complete, we mean that if we consider a subset of preference and weight statements, the decision remains unchanged regardless of how this subset is completed. For simplicity, we have considered a cost function with different properties (neutrality, monotony, additivity), in which we try to capture the simplicity of the statement as the easiness for the user to understand it. Let us consider the example again.

> **Example 4.4. (Ex. 4.1 Cont.)**
>
> A not complete explanation (it does not provide enough evidence that a is preferred over c):
>
> $$E_1 = [1, 2 : a \succ d, e, f], [1, 3 : a \succ b], [2, 3 : a \succ c]$$
>
> A complete explanation:
>
> $$E_2 : [1 : a \succ e, d, b, f], [2 : a \succ f, e, d, c], [3 : a \succ b, c, d], [4 : a \succ c, e], [5 : a \succ c]$$

In the previous example, one can note that E_2 is certainly not minimal since (for instance) the same explanation without the last statement is also a complete explanation whose cost is certainly lower (by monotonicity of the cost function). Now if the cost function is sub-additive, then a minimal explanation cannot contain (for instance) both $[1, 2 : a \succ d, e]$ and $[1, 2 : a \succ f]$. This is so because then it would be possible to factor these statements as $[1, 2 : a \succ d, e, f]$, all other things being equal, to obtain a new explanation with a lower cost.

Among others, an interesting result from this work is that minimal explanations are free of negative statements, and neutral ones can be ignored. We proposed a polynomial computation of a minimal element of the explanation with the basic preference statements. However, the additional expressive power provided by the factored statements comes at a price when we want to compute minimal explanation, as it is stated by Proposition 4.1.

Proposition 4.1. *([Labreuche et al., 2011]) Deciding if (using factored statements) there exists an explanation of cost at most k is NP-complete. This holds even if criteria are unweighted and if the cost of any statement is constant.*

The previous result shows that no efficient algorithm can determine minimal explanations when the cost function implies minimizing the number of factored statements (unless P=NP). This is true unless we restrict to specific classes of cost functions; thus, the problem may turn out to be easy. In this work, we discussed two cases. First, when the cost function is super-additive, it is sufficient to look for basic statements. Second, when it is sub-additive, an idea could be to restrict the attention to statements which exhibit winning coalitions. In this case, the problem can be turned into a weighted set packing, for which the direct Integer Linear Program formulation would be sufficient for a reasonable size of options and criteria sets. Finally, enforcing a complete explanation implies a relatively large number of items in the explanation. However, in most cases, factored statements allow for obtaining short explanations.

4.2.1.2   Explanations when PI is incomplete

A decision model is specified from some PI provided by the decision-maker during an interview, related to comparing the options on each criterion and the weights of the criteria. However, the PI is insufficient to specify the model most of the time. In particular, some options may be incomparable on some criteria for the decision-maker. Moreover, the elicitation process (see Figure 2.1) will not result in a single value of the weight vector but rather in a set of vectors that are compatible with the PI [Greco et al., 2010]. Then, an option a is said to be *necessarily* preferred to another one b if the first option is preferred to the second one (noted $a \succ b$) for all weight vectors that are compatible with the PI and for all ordering of the options on the criteria that are compatible with the PI [Greco et al., 2010].

Considering this incompleteness of PI, we investigated the question of searching and defining a simple explanation for a recommended choice. Thus, we are looking to justify that a given option is a weighted Condorcet winner (WCW), i.e. this option is necessarily preferred to each other option, whatever the weight vector compatible with the PI. However, instead of the first case, if the WCW does not exist, we will consider the Smith set [Fishburn, 1977]. It is the smallest set of alternatives such that all the elements in this set beat the elements outside it. When the WCW exists, the Smith set is reduced to the WCW.

As in the previous case, we need information regarding the ranking of options and the relative strength of coalitions of criteria. For illustration, let us take Example 4.5, where option a is the WCW and the unique dominating option (that beats all the other options).

Example 4.5. [Labreuche et al., 2012]

There are 7 options $\{a, b, c, d, e, f, g\}$ and 4 criteria $\{1, 2, 3, 4\}$. The partial orderings (noted $\succ_1, \succ_2, \succ_3, \succ_4$) of options over the 4 criteria are depicted in Figure 4.1. The PI regarding the importance of the criteria is composed of the following three statements:

- 1 together with 3 are more important than 2 and 4 together;
- 2 and 3 together are more important than criterion 1 taken alone;
- 4 is more important than criteria 2 and 3.
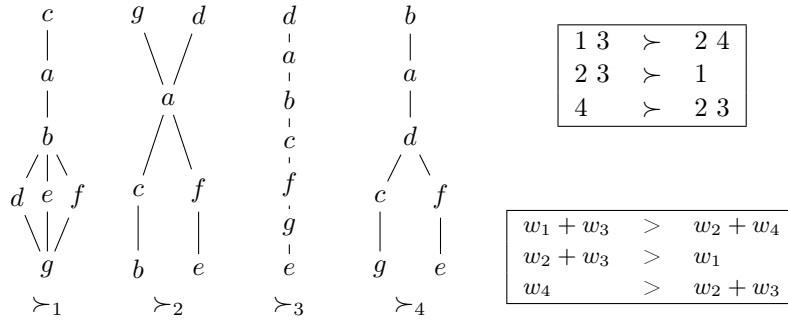
Figure 4.1: Partial preferences $\succ_1, \succ_2, \succ_3, \succ_4$ over the criteria 1,2,3,4.

Now, the "technical" reasons why a is the WCW are depicted in Ex. 4.6.

Example 4.6. (Ex.4.5. Cont.)

(i) a dominates e and f on all criteria,

(ii) $a \succ b$ because it is better on the coalition 123,

(iii) $a \succ d$ because it is better on the coalition 14,

(iv) $a \succ g$ because it is better on the coalition 134,

(v) $a \succ c$ because it is better on the coalition 234.

First, to express such explanations, we need two types of statements. First, a set of *preference statements* (noted $\mathcal{S}$) (the comparison of an option over another one on a given criterion). Second, a set of *comparative statements* (noted $\mathcal{V}$) (stating the importance among two disjoint subsets of criteria). Therefore a PI is a pair $\langle S, V \rangle$ with $S \subset \mathcal{S}$ and $V \subset \mathcal{V}$.

It is also important to note that expressing a comparative statement (e.g. $13 \succ 24$

) amounts to expressing a constraint ($w_1 + w_2 > w_2 + w_4$) on the feasible region of the feasible weight vector attached to the criteria (see Example 4.5). Moreover, the information provided by the decision-maker is supposed to be "rational". Specifically, S constitutes a partial order (reflexive, antisymmetric, transitive, but not complete), and V is assumed to be consistent[4].

Example 4.7. (Ex.4.5. Cont.)

Given the PI of Example. 4.5, $V = \{[13 \succ 24], [23 \succ 1], [4 \succ 23]\}$.
We have for instance, $[c \succ_1 d] \in S$, $[b \succ_2 a] \notin S$, and $\langle 0.2, 0.1, 0.15, 0.55 \rangle$ is not a compatible vector of weights (violation of the first constraint).

Second, let us analyze the reasons depicted in Example 4.6. One can notice that these reasons vary in terms of the effort required to understand them: (i) is trivial, and (ii), (iii) and (iv) are reinforcement of some statements of the PI. For instance, (ii) quickly follows from the fact that 1 and 3 are already more important than 2 and 4. On the other hand, the underlying justification for (v) is more complex. *How to deduce from the PI the statement that coalition 34 beats coalition 12?* In other terms, imagine that in the ordering $\succ_2$, c is now preferred to a. Is it true that $a \succ c$ because it is supported by the coalition 34?

Therefore, it appears that dominated option can be partitioned into different classes, capturing the fact that some of them are *obviously* dominated, some are *clearly dominated*, while others are close to a tie with some elements of the dominating set. These different situations will be called by: *unanimous*, *large majority* and *weak majority*. The first case does not require any specific explanation. The second is a clear-cut situation that may need only a rough explanation. In the last case, the decision is unclear, and a detailed explanation is required. In the following, we will focus our development mainly on this case (for more details, see [Labreuche et al., 2012]).

To construct the explanation for the *weak majority* case, we can try to apply the approach presented in Section 4.2.1.1, where providing an explanation amounts to simplifying the PI provided by the decision-maker (here, the pair $\langle S, V \rangle$)) as long as the same decision holds. However, as we shall see with Example 4.8 it is not enough to provide a convincing explanation.

---

[4]In fact, many works with explanation in AI address the problem of exhibiting subsets of constraints provoking an inconsistency, see, e.g. [Junker, 2004]

Example 4.8.

Consider five criteria and four options a, b, c, d. Assume that
$V = \{[1 \succ 23], [34 \succ 15], [2 \succ 5]\}$ and $S = \{[a \succ_1 b], [a \succ_4 b], [a \succ_5 b], [a \succ_2 c], [a \succ_3 c], [a \succ_4 c], [a \succ_1 d], [a \succ_3 d], [a \succ_4 d], [b \succ_3 d]\}$.
Let $V' = \{[1 \succ 23], [34 \succ 15]\}$ and $S' = S \setminus \{[b \succ_3 d]\}$.

Indeed, in Example 4.8, the pair $\langle S', V' \rangle$ is the minimal complete explanation, in the sense of set inclusion, justifying that a is the WCW. For instance, in the produced explanation, we have "$a \succ d$ because a is better than d on the coalition 134". However, from only V', it is unclear why 134 is a winning coalition! Nevertheless, it clearly follows from $[13 \succ 25]$. Hence reduction over V does not simplify the explanation! In other words, we observe that to support a WCW; we may use new comparative statements (e.g. 134) deduced from the set of comparative statements of the PI. Therefore, explaining a WCW in this situation amount to not only proving that an option is certainly a WCW but also being able to explain why the supporting coalition is indeed a winning one.

Thus, to construct a simple and complete explanation when the PI is incomplete, we need two components, (i) explaining why an option is a WCW (we build S' by simplifying S, in the sense of set inclusion) and (ii) explaining why the supporting coalition is a winning one. For the latter we characterized an operator $\mathtt{cl}$ such that $\mathtt{cl}(V)$ is the set of comparative statements that can be deduced from V. This characterization shows that all comparative statements deduced from V result from a linear combination (with integer coefficients) of the constraints in V and of the constraints on the sign of the weights (we rely on the Farkas Lemma for this characterization). To illustrate this idea of linear combinations, consider Example 4.9.

Example 4.9.(Ex. 4.8. Cont.)

(i) $[14 \succ 23] \in \mathtt{cl}(V)$ follows from $[1 \succ 23]$, by monotonicity.

(ii) $[4 \succ 25] \in \mathtt{cl}(V)$ follows from $[1 \succ 23]$ and $[34 \succ 15]$, because

$$
\begin{array}{rcl}
w_1 & > & w_2 + w_3 \\
+ \quad w_3 + w_4 & > & w_1 + w_5 \\
\hline
= \quad w_1 + w_3 + w_4 & > & w_1 + w_2 + w_3 + w_5 \\
= \quad \cancel{w_1} + \cancel{w_3} + w_4 & > & \cancel{w_1} + w_2 + \cancel{w_3} + w_5
\end{array}
$$

Moreover, by examining the elements belonging to $\mathtt{cl}(V)$, we noticed that it was possible to organize the latter into four nested sets. These sets correspond to difficulty classes in justifying an element from V. More precisely, we can distinguish, from the

lowest to the highest complexity, comparative statement: (i) $\mathtt{cl}_0(V)$ contained directly in the PI (no underlying complexity for the user, e.g. $[23 \succ 1]$ in Ex. 4.5), (ii) $\mathtt{cl}_1(V)$ that can be deduced from V only using monotonicity (e.g. $[4 \succ 3]$), (iii) $\mathtt{cl}_2(V)$ that can be deduced from V only using summation and monotonicity conditions (e.g. $[4 \succ 1]$), and (iv) $\mathtt{cl}_3(V)$ that are in $\mathtt{cl}(V)$ (e.g. $[34 \succ 21]$). Therefore, the target is to construct an explanation, when it is possible, with the smallest number of the last category and to build on the less complex ones. In the end, an efficient algorithm is provided to compute the minimal explanation by considering mainly three steps: determining the comparative statements of the different complexity classes ($\mathtt{cl}_j(V)$, $j \in \{1, 2, 3\}$), identifying all the preference statements ($S' \subset S$) that justify the WCW such that $\mathcal{V}(S') \subset \mathtt{cl}(V)$, and finally determining elements of S' such that the explanation is minimal in the sense of the order that depicts the complexity of understanding why a set of comparative statement derives from V.

To conclude, a distinctive feature of our approach lies in the decision model, taken together with the fact that the PI may be largely incomplete. In this context, the precise weights attached to attributes cannot be exhibited, and the challenge is to provide convincing (complete) explanations despite this constraint.

4.2.2   Explaining pairwise comparisons

We explore the problem of providing explanations for pairwise comparisons based on an underlying additive model. We follow a step-wise approach and provide explanations that take the form of a sequence of preference statements. The explanations we aim for are thus *contrastive*, in the sense that the decision to be explained compares two alternatives, and *exact* (as opposed to *heuristic*) in the sense that we provide guarantees that the explanation produced is correct concerning the underlying model. It is also common to distinguish between *local* explanations (when they focus on a specific recommendation) and *global* explanations (when they deal with the model in general): our approach is globally faithful to the model and locally relevant to the pairwise comparison to be explained. Let us consider the following illustrative example to make things more concrete.

Example 4.10. (Motivating Example)

We consider seven abstract criteria ($\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g}$), each one described on bi-levels scales, which facilitate the symbolic representation of alternatives (*e.g.* hotels). Each alternative can be represented as its evaluation vector ($s_1 = (\mathtt{✗}, \mathtt{✗}, \mathtt{✓}, \mathtt{✓}, \mathtt{✓}, \mathtt{✓}, \mathtt{✓})$) or more succinctly by the subset of criteria on which it is evaluated positively ($s_1 = \{\mathtt{cdefg}\}$). Moreover, for each criterion, the value

symbolized by ✓ is more desirable than the value symbolized by ✗ (*e.g.* breakfast included is better than not).

	a	b	c	d	e	f	g
s_1	✗	✗	✓	✓	✓	✓	✓
s_2	✓	✗	✗	✓	✗	✗	✗

The aggregation of criteria is done using an additive score function, assigning weights to the different criteria. The function is as follows:

$$w = \langle 128, 126, 77, 59, 52, 41, 37 \rangle$$

For example, the score of s_1 is thus equal to $score(s_1) = 77+59+52+41+37 = 276$ while that of s_2 is: $score(s_2) = 128 + 59 = 187$. It is also useful to encode the comparison of two alternatives as a vector $\{-1, 0, +1\}^n$ of arguments in favour (PRO) or against (CON) s_1, or neutral (NEU). In our example, PRO $= \{c, e, f, g\}$, CON $= \{a\}$, while NEU $= \{b, d\}$

Explanations can take many different forms. We list different possible explanations for the fact that s_1 is preferred to s_2:

(i) the first approach (*model disclosure*) could be to provide the full score calculation for both options, as illustrated above. However, noticing that d is a neutral argument satisfied both by s_1 and s_2, we could omit it and provide the summation of PRO arguments vs CON arguments.

(ii) the *counter-factual* approach seeks minimal modification in the input that would change the outcome. For instance, we could state that, if s_2 had satisfied b, s_2 would instead have been recommended over s_1. Or (affecting the other alternative this time), if s_1 had not satisfied cd.

(iii) Following a *prime implicant* approach, we could produce sufficient arguments to explain the decision. In our case, two possible explanations could be given: (1) given that bd are neutral arguments, the PRO arguments cef are sufficient to overcome any set of CON arguments. In particular, this shows that the decision would remain the same even if g was a CON argument. Moreover, (2) given that b is a neutral argument, the PRO arguments cefg are sufficient to overcome any set of CON arguments. In particular, this shows that the decision would remain the same even if d was a CON argument.

(iv) following a *step-wise* approach, we could exhibit a collection of statements aiming

at proving the decision. For instance, we could state that `cdefg` is preferred over `ac`, and that `ac` is preferred over `ad`, so that our conclusion should hold, following a *transitive* reasoning. Alternatively, using a different logic, we could state that `cd` is preferred over `a`, while `efg` is preferred over `d`, which altogether justifies our decision.

Our main idea is to break down the recommendation into "simple" statements presented to the explainee. The whole sequence of statements should formally support the recommendation. We build on the notion of *argument schemes*, that is, an operator tying a sequence of statements called premise, satisfying some conditions, into another statement called the conclusion [Walton, 1996]. As we deal with preferences, argument schemes are ways of deriving new preferences from previously established ones. More precisely, we consider a set of items $[m]$, and we abstractly refer to *states*, as subsets of items, i.e. elements of $2^{[m]}$. A *comparative statement* is a pair of states $(A, B) \in 2^{[m]} \times 2^{[m]}$, interpreted as a preference statement–'A is preferred to B'. Thus, our schemes operate on the same set of premises – finite sequences of comparative statements, represented as bracketed lists – and the same set of conclusions. We shall denote an arbitrary scheme s as:

$$[(A_1, B_1), \ldots, (A_k, B_k)] \xrightarrow{s} (A, B)$$

More precisely, we propose to develop a principle-based and cognitively bounded model of step-wise explanations. Our view of explanations as cognitively bounded deductive proofs is reminiscent of the *bounded proof systems* proposed in the context of description logic [Horridge et al., 2013; Engström and Abdul Rahim Nizamani, 2014]. Also, a similar step-wise approach has been studied in the context of constraint satisfaction problems [Bogaerts et al., 2021]. Finally, a close setting the one of explanations based on axioms have been advocated in computational social choice [Cailloux and Endriss, 2016; Procaccia, 2019]. In particular, the recent work of [Boixel et al., 2022] also exploits axioms studied in voting theory to produce explanations for collective decisions but applied to a different setting (voting) and using different proof techniques (tableau methods).

As our example illustrates, there can be different 'logic' at play when combining statements. To account for that we proposed a number of *argument schemes* in the context of a pairwise comparison based on a weighted sum model (see Figure 4.2, where an arrow from $scheme_1$ to $scheme_2$ denotes that all instances satisfying $scheme_2$ also satisfy $scheme_1$, but not the converse.).

By principle-based approach, we mean that each scheme is attached to a number of well-understood properties of the underlying decision model (see Table 4.2) that we make explicit. Obviously, an additive preference satisfies both the transitive and cancel-
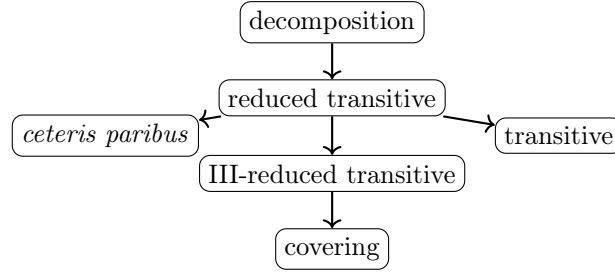
Figure 4.2: Relationships between argument schemes

lation properties. The resulting calculus is provably correct. By cognitively bounded, we mean that our statements will be constrained to remain easy to grasp by the explainee. This has the consequence of making the resulting calculus *not* complete. However, we explore this issue in detail and provide several elements showing that our approach is satisfactory in terms of empirical completeness (see the discussion at the end of this section).

| | | Requirements for |
Scheme	Properties	correctness
decomposition	commutative	additive
reduced transitive		transitive + cancellation
III-red. transitive	III	transitive + cancellation
covering	commutative, III	transitive + cancellation
transitive		transitive
ceteris paribus		cancellation

Table 4.2: Structural properties of the reasoning schemes.

Moreover, we want an explanation to be "easy to process" by the explainee. Thus, it requires specifying the relative difficulty of a premise and a conclusion. We introduce a specific model allowing us to derive the relative difficulty of statements, where this difficulty is purely syntactic and directly results from the number of items involved in the comparative statement. Thus, we define what we call *difficulty classes* of comparative statements by putting upper bounds on the difficulty: for all integers p, q from 0 to m, let $\Delta(p, q) = \{(A, B) \in 2^{[m]} \times 2^{[m]} : |A| \leq p, |B| \leq q\}$. These classes specify the set of atomic elements considered self-evident and legit to be used as steps of an explanation for the considered explainee. In the context of explaining preferences between a subset of desirable items, some values of the pair (p, q) are of specific interest: $\Delta(m, m)$ are unrestricted statements; comparative statements in $\Delta(m, 0)$ represent Pareto dominance statements; comparative statements in $\Delta(1, 1)$ can be interpreted as *swaps* [Hammond et al., 1998], representing the exchange of one criterion against another; those in $\Delta(1, m)$ or in $\Delta(m, 1)$ represent a single item stronger or weaker than a subset of others, respectively considered as a pro or a con argument. For instance,

in the context of hotel comparisons, an argument in $\Delta(1,1)$ could be *"we prefer to have free breakfast then free wifi access"*. An argument in $\Delta(1,2)$ could be *"We prefer to have a swimming pool than free breakfast and wifi"*. To appreciate how difficult it can be to interpret higher-order arguments, consider arguments in $\Delta(2,2)$. These could correspond to *"free breakfast and wifi access are preferable to having a swimming pool and being close to the city centre"*. We investigate how restraining explanation to use these classes of simple statements affects the production of an explanation. Some insights later in this section.

To give an overview of this work, we propose briefly describing only two examples of schemes, namely the decomposition scheme [Belahcene et al., 2019] and the covering scheme [Belahcene et al., 2017a]. Moreover, when it is possible and not confusing, we propose skipping the technical details to give only a high-level overview through illustrative examples. For more details, we refer the reader to [Amoussou et al., 2022] We draw the attention of the reader that when we have only transitive schemes and dominance, we are in the situation of [Labreuche et al., 2012] (see 4.2.1.2).

The decomposition Scheme.   Introduced in [Belahcene et al., 2019] and implementing cancellation properties of higher order [Krantz et al., 1971; Wakker, 1989], the decomposition scheme aims at leveraging the assumed additive property of the preference relation[5]. When a preference is additive, preference statements translate into linear comparisons that can be summed up. Then, the scores of items appearing on both sides cancel out, sometimes allowing to derive new comparisons. In other words, this scheme operates by interpreting a Farkas certificate as sets of arguments, pros and cons for a preference statement, then carving the desired conclusion through a cancellative property. Consider Example 4.11 for illustration.

Example 4.11. (Decomposition Scheme)

Consider the following decomposition scheme:

$$[(\mathtt{bc}, \mathtt{de}), (\mathtt{efg}, \mathtt{ac})] \xrightarrow{dec} (\mathtt{bfg}, \mathtt{ad})$$

Assuming that the preference $\succsim$ is additive, and that both $\mathtt{bc} \succsim \mathtt{de}$ and $\mathtt{efg} \succsim \mathtt{ac}$. From the first comparison, we deduce that $\omega_\mathtt{b} + \omega_\mathtt{c} \geq \omega_\mathtt{d} + \omega_\mathtt{e}$; from the second that $\omega_\mathtt{e} + \omega_\mathtt{f} + \omega_\mathtt{g} \geq \omega_\mathtt{a} + \omega_\mathtt{c}$. By summation, we derive $\omega_\mathtt{e} + \omega_\mathtt{f} + \omega_\mathtt{g} + \omega_\mathtt{b} + \omega_\mathtt{c} \geq \omega_\mathtt{d} + \omega_\mathtt{e} + \omega_\mathtt{a} + \omega_\mathtt{c}$.

---

[5]This decomposition scheme is less general than the so-called *syntactic cancellative* described in [Belahcene et al., 2019], as it does not allow for repetition of the conclusion. This has been shown to reduce expressiveness.

Then, as it is illustrated in the following by cancelling ω_e and ω_c on both sides (this is actually an instance of *second order cancellation*, because it is performed across two comparative statements), we obtain $\omega_f + \omega_g + \omega_b \geq \omega_d + \omega_a$, hence bfg $\succ_\omega$ ad.

$$
\frac{b \quad \cancel{c} \qquad\qquad \succ \qquad\qquad d \quad \cancel{e}}{\cancel{e} \quad f \quad g \quad \succ \quad a \quad \cancel{c}}
$$
$$
b \qquad\qquad f \quad g \quad \succ \quad a \qquad d
$$

The Covering Scheme. The covering scheme particularizes both the reduced transitive and decomposition schemes (see Figure 4.2). In this scheme a list of comparative statements $[(A_1, B_1), \ldots, (A_k, B_k)]$ supports a conclusion (A, B) if, and only if, the *pros* $A_1, \ldots, A_k$ partition $A \setminus B$ and the *cons* $B_1, \ldots, B_k$ partition $B \setminus A$.

Example 4.12. (Covering Scheme)

Consider the conclusion: $(\mathtt{bfg}, \mathtt{cde})$. The premise $[(\mathtt{fg}, \mathtt{c}), (\mathtt{b}, \mathtt{de})]$ constitute a covering scheme:

$$
[(\mathtt{fg}, \mathtt{c}), (\mathtt{b}, \mathtt{de})] \xrightarrow{cov} (\mathtt{bfg}, \mathtt{cde})
$$

On the one hand, the scheme formalizes a proof, articulating transitive (tr) and *ceteris paribus (cp)* derivations that can be presented to the explainee as a diagram, such as in Example 4.13, or narratively such as in Figure 4.4 (for hotel comparisons for instance). On the other hand, the premises can be understood as grouping some cons with some stronger pros so as to "cover" the cons and can be presented visually to the explainee, such as in Figure 4.3.

Example 4.13 (Three representations of the Covering Scheme).

$$
\left.\begin{array}{l} \mathtt{fg} \succ \mathtt{c} \xrightarrow{cp} \mathtt{bfg} \succ \mathtt{bc} \\ \mathtt{b} \succ \mathtt{de} \xrightarrow{cp} \mathtt{bc} \succ \mathtt{cde} \end{array}\right\} \xrightarrow{tr} \mathtt{bfg} \succ \mathtt{cde}
$$

Covering Scheme: proof diagram of Ex. 4.12
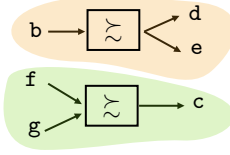
Figure 4.3: Covering scheme: a visual representation of Ex. 4.12

*"As, all other things being equal, having free breakfast and wifi access is preferred to having a swimming pool (**fg**, **c**), and being close to the city is preferred than having a sports hall and a low tourist tax (**b**, **de**), we get that (**bfg**, **cde**)"*

Figure 4.4: Covering scheme: a narrative representation of Ex. 4.12

We have investigated the relative expressiveness and computational complexity of explaining with the reduced transitive and the covering schemes, together with the choice of atomically simple statements. It results that without any restriction on the set of atomic statements ($\Delta(m,m)$), it is difficult (NP-hard) to decide whether an explanation exists with these schemes. Regarding the other schemes, while *ceteris paribus* scheme is easy, we conjuncture the complexity of decomposition and III-reduced transitive to be intractable.

Now, when we put syntactic restrictions on the sets of atomic elements used, $\Delta(1,1)$, $\Delta(1,m)$, $\Delta(m,1)$, among the results, we state that the covering scheme is transitive. A similar result has been identified in [Belahcene et al., 2017a] with the restricted $\Delta(1,1)$, where we have proposed an explanation mechanism that produces an explanation under the form of a chain of *transitive* statements, restricted to the expression of trade-offs between at most m points of view. This approach takes its inspiration in the *even swaps* interactive elicitation mechanism [Hammond et al., 1998], then turns it upon its head – assuming the model is known rather than trying to build it and expressing mere preference statements rather than asking cardinal information making an alternative indifferent to another. Thanks to the characterization of the necessary preference relation [Belahcene et al., 2017a], we showed that, with the additional assumption of using only two levels on every criterion when collecting preferential information, sequences of preference swaps of order at most two, $\Delta(1,1)$, have a term by a term structure that ensures they have a short length (at most half the number of criteria) and they can be efficiently computed. However, when $m \geq 2$, the problem is difficult. Moreover, although the different schemes may correspond to alternative explanation strategies, we specifically advocate using the covering scheme

because it meets some desirable properties of explanations. Therefore, we studied the empirical completeness of atomic statements $(\Delta(1,m),\ \Delta(m,1))$ using the covering scheme. With this scheme, we can say that a significant majority of the pairs to explain are explainable. For example, for $m = 6$, more than 3 pairs out of 4 are explainable regardless of the *additive linear order* considered.

Finally, we note that in [Belahcene et al., 2017a], the explanation of pairwise comparison is constructed for a necessary preference relation [Greco et al., 2010], which makes minimal assumptions while handling a collection of compatible utility functions, which are impossible to exhibit to an end-user. The problem with such an explanation is that it is not always easy to construct it; even in some situations does not exist. Therefore, in [Amoussou et al., 2020; Amoussou, (in progress)], we proposed alleviating some of the preference-swaps explanation constraints to arrive at what we call a *mixed explanation*, where the computation of its components is done through the resolution of a Mixed Integer Linear Program. These elements belong to both necessary and possible preference swaps. The possible swaps correspond to a subset of additive utility functions compatible with the preference information. One note that providing a sequence composed of solely necessary swaps guarantees that the recipient of the explanation will accept and validate each swap without any doubt, which is not the case with the possible swaps. However, we believe that using possible swaps offers a way to collect more additional preference information (valuable in a preference elicitation process) and thus enrich both the preference information and the necessary relation. The idea is to rely on the statements involved in the explanation to allow the explainee to accept or contradict these statements and thus benefit from this feedback to enrich the learning task and validate the model. Indeed, we think that in a decision support situation, at a given moment, the initiative should be left to the user to express an opinion when confronted with the explanation. This idea is discussed more in detail in Chapter 5.

4.2.3   Explaining an assignment

This section is devoted to describing how the theoretical and algorithmic tools described in Section 3.3.1.2 in order to assess the feasibility of the inverse NCS problem can be used to support a decision process. The technical details of this work can be found in [Belahcene et al., 2018b].

More precisely, we address the situation described in Example 4.14 where a committee meets to decide upon sorting several candidates into two categories (e.g. candidates to accept or not, projects to fund or not). The committee applies a public decision process; the outcomes are also public. However, the details of the votes are sensitive and should not be made available. To what extent can we make the committee accountable for its decisions?

We are interested in a general sorting model where candidates are sorted by a jury N. Each juror ⚥ $\in N$ expresses binary judgements [Laslier and Sanver, 2010], and candidates are sorted either to the GOOD or the BAD category, depending on the fact that the coalitions of jurors supporting this sorting are strong enough, or not, to win the decision of the jury.

Example 4.14.

We consider a situation with six candidates $\mathbb{X} := \{a, b, c, d, e, f\}$, assessed by a jury composed of five jurors $N := \{$⚥^{1}, ⚥^{2}, ⚥^{3}, ⚥^{4}, ⚥$^5\}$ with the following preferences

$$
\begin{aligned}
⚥^1 : \quad & a \succ_1 b \succ_1 f \succ_1 e \succ_1 c \succ_1 d \\
⚥^2 : \quad & e \succ_2 b \succ_2 c \succ_2 d \succ_2 a \succ_2 f \\
⚥^3 : \quad & f \succ_3 a \succ_3 b \succ_3 d \succ_3 e \succ_3 c \\
⚥^4 : \quad & d \succ_4 a \succ_4 c \succ_4 e \succ_4 f \succ_4 b \\
⚥^5 : \quad & c \succ_5 e \succ_5 b \succ_5 f \succ_5 d \succ_5 a
\end{aligned}
$$

Adopting the primitives of the Non-Compensatory Sorting model: candidates are *alternatives*, jurors are *points of view*, and we are considering two *categories* { BAD $\prec$ GOOD }. For the NCS model to correctly describe the situation, the decision process needs to be bounded by some assumptions of rationality.

- *Static individual stances.* From the personal point of view of each juror, alternatives should be completely preordered by preference. This precludes any incomparability between candidates nor dynamics in how each juror appreciates the candidates.

- *Indivividual consistency between preferences and vote.* Each juror ⚥ $\in N$ is allowed to express only a binary judgment on each candidate $x \in \mathbb{X}$, which is either 'approved according to ⚥' or not. The approved subset of candidates $\mathcal{A}_⚥ \subseteq \mathbb{X}$ should be an upset for the preference relation $\succsim_⚥$. Hence, there is no pair of candidates $x, x' \in \mathbb{X}$ where x is preferred to x' w.r.t. $\succsim_⚥$, x' is approved by ⚥ but not x.

- *Static collective stance.* The set of winning coalitions should remain constant during the whole decision process. This can be seen as a requirement for the process to be unbiased.

- *Consistent collective stance.* The set of sufficient coalitions $\mathcal{S} \subseteq \mathcal{P}(N)$ should be an upset for inclusion. Hence, if a coalition is sufficient, any superset of this

coalition is also sufficient (and if a coalition is insufficient, any subset of it is also insufficient).

- *Latent coalition powers.* The set of sufficient coalitions is not assumed to have any particular structure besides being an upset.

Example 4.15.

Suppose the approved sets are as follows:
$\mathcal{A}_{\mathbf{Q}^1} := \{a, b, f\}, \mathcal{A}_{\mathbf{Q}^2} := \{e, b, c\}, \mathcal{A}_{\mathbf{Q}^3} := \{f, a, b\}, \mathcal{A}_{\mathbf{Q}^4} := \{d, a, c\},$
$\mathcal{A}_{\mathbf{Q}^5} := \{c, e, b\}$, corresponding to the three best alternatives according to the respective points of view (3-approval).

Suppose also the points of view are aggregated according to the simple majority rule, i.e. $B \in \mathcal{S} \iff |B| \geq 3$. Then, the corresponding non-compensatory model assigns a, b, c to the GOOD category, and d, e, f to the BAD one. Hence, $\alpha := \{(a, \text{GOOD}), (b, \text{GOOD}), (c, \text{GOOD}), (d, \text{BAD}), (e, \text{BAD}), (f, \text{BAD})\}$.

We note the same assignment α can be obtained with different sorting parameters, e.g. approved sets $\mathcal{A}'_{\mathbf{Q}^1} := \{a, b, f\}, \mathcal{A}'_{\mathbf{Q}^2} := \{e, b, c, d, a\}, \mathcal{A}'_{\mathbf{Q}^3} := \{\}, \mathcal{A}'_{\mathbf{Q}^4} := \{d, a, c\}, \mathcal{A}'_{\mathbf{Q}^5} := \{c\}$ and sufficient coalitions $\mathcal{S}'$ containing the coalitions $\{\mathbf{Q}^1, \mathbf{Q}^2\}, \{\mathbf{Q}^5\}$ and their supersets.

While the jury as a whole has the power to take decisions, we consider a situation where it has to account for its decisions. This requirement may take several forms, and we focus our attention on two specific demands:

- **Procedural regularity.** Kroll et al. [2017] puts forward that a baseline requirement for accountable decision-making–and, therefore, a key governance principle enshrined in law and public policy in many societies[6]–is *procedural regularity*: each participant will know that the same procedure was applied to her and that the procedure was not designed in a way that disadvantages her specifically.

- **Contestability.** An attractive normative principle [Pettit, 1997, 2000] is contestability: a democratic institutional arrangement should be such that citizens can effectively challenge public decisions. The control of the governed on the government is generally two-dimensional: electoral and contestatory. For reasons of practical feasibility, administrative decisions are typically under contestatory control. In this context, a candidate (supposedly) unsatisfied with the outcome

---

[6]E.g. by the Fourteenth Amendment in the USA.

of the process regarding his own classification could challenge the committee and asks for a justification.

A typical way to address *procedural regularity* is to require *transparency* and let an independent audit agency access all the available information. Transparency could also be an adequate answer to *contestability*, provided the decision rule is *interpretable* (comprehensible by the persons that need to–here, the contestant). In the context of jury decisions, transparency is out of the question, as it suffers from several drawbacks:

Sensitive information.   In this setting, the 'details of the votes' cover two aspects: (i) the approval of jurors at the individual level; and (ii) the winning coalitions at the jury level.
 These details might be worth considering as sensitive information for several reasons:

- Protecting the jurors from external pressure, including threats or retaliation.

- Protecting the jury and jurors from internal pressure: maybe the approval procedure should be made with secret ballots. Maybe revealing the actual balance of power inside the jury could exacerbate tensions.

- The details of the approval of each candidate might be considered personal information belonging to each candidate and should not be disclosed to third parties.

- Revealing dissension among the jurors might weaken the jury's authority.

- Revealing the decision rule, or publishing much information about it, would create a feedback effect with some candidates adopting a strategic behavior to game the output.

Complexity   Leaving the burden of proof on the shoulders of the audit agency, or worse, of a lone plaintiff, may be too demanding. At the same time, it requires access to much information–possibly the preferences and the assignment of the whole set of candidates–and to solve complex combinatorial problems that scale poorly with the number of candidates. Indeed, we have shown that the Inv-NCS problem is NP-hard [Belahcene et al., 2018b].

In what follows, we describe how to address the procedural regularity and the contestability requirements while paying attention to disclosing as little information as necessary and providing comprehensible explanations by their recipient.

Addressing overall *Procedural regularity* with Inv-NCS.   The question addressed here is how observers can be assured that each sorting decision was made according to the same procedure. Because of this demand, what needs to be proven

is that α is a positive instance for the Inv-NCS problem (see Section 3.2), i.e. the assignment α is a *possible* outcome for NCS, given the preferences of the jurors over the candidates.

Should the burden of proof be left to the auditor, the audit procedure could require either:

i) full disclosure of the preference profile $\langle(\mathbb{X}, \succsim_i)\rangle_{i \in N}$, and the auditor solving the NP-hard Inv-NCS problem, e.g. using a SAT solver and either of the formulations Φ_α^C or Φ_α^P described in Chapter 3, or

ii) full disclosure of the approved sets $\langle \mathcal{A}_i \rangle_{i \in N}$, and the auditor solving the polynomial-time problem Inv-NCS with fixed accepted sets problem as described in Chapter 3, Equation 3.3.

Note that the entire disclosure of the decision rule is not an option. It would require revealing the entire parameter specifying the NCS model and, in particular, the provision of the set of sufficient coalitions. This is impossible, as the *ground truth*, i.e. the rule deciding which coalition is sufficient, is oral at best and most likely implicit. We consider the jury has black-box access to it, and the external auditor can only guess the contours of this rule through indirect evidence. It is likely that the investigations made by the audit agency reveal *possible parameters* that do not correspond to the ground truth. If we consider putting the burden of proof on the committee, a third option can be engineered. We propose to leverage Theorem 3.1 to compute and provide a certificate of feasibility for Inv-NCS(α) that involves the disclosure of less information, as illustrated below:

Example 4.16. (Ex. 4.15 Cont.)

If the approved sets of the committee are $\mathcal{A}_{⚥^1}, \ldots, \mathcal{A}_{⚥^5}$, then it needs to disclose some information concerning three points of view in order to prove the assignment α is consistent with an approval procedure, e.g. :

- according to the first juror $⚥^1$:

  - b is approved;
  - a is preferred to b;
  - e is not approved;
  - e is preferred to d;

  therefore, the procedure is able to positively discriminate a, b from d, e;

- according to the second juror $⚥^2$:

  - c is approved;

- b is preferred to c;
- d is not approved;
- d is preferred to f;

therefore, the procedure is able to positively discriminate b, c from d, f;

- according to ⚥^{4}:

  - c is approved;
  - a is preferred to c;
  - e is not approved;
  - e is preferred to f;

therefore, the procedure is able to positively discriminate a, c from e, f.

The following table summarizes the jurors known to discriminate each pair:

		BAD		
		d	e	f
	a	⚥^{1}	⚥^{1}	⚥^{4}
GOOD	b	⚥^{1}	⚥^{1}	⚥^{2}
	c	⚥^{2}	⚥^{4}	⚥^{2}

As every pair in $\{a, b, c\} \times \{d, e, f\}$ is positively discriminated by at least one member of the jury, the procedure is regular: there is, for each juror individually and for the jury, collectively, a way of proceeding accordingly to the principles exposed at the beginning of this section, and deem $\{a, b, c\}$ GOOD and $\{d, e, f\}$ BAD .

This manner of arguing that a given assignment is indeed a possible outcome of an approval sorting procedure has been formalized into an argument scheme (described formally in [Belahcene et al., 2018b] and illustrated in Example 4.17.

Example 4.17.

The explanations given in Example 4.16 are as follows: $\langle (⚥^1, b, \{a, b\}, e, \{d, e\}),$ $(⚥^2, c, \{b, c\}, d, \{d, f\}), (⚥^4, c, \{a, c\}, e, \{e, f\}) \rangle$

- according to the first point of view, b is approved (and so is a which is better than b) whereas e is not (and neither is d which is worse than e),

- according to the second point of view, c is approved (and so is b which is better than c) whereas d is not (and neither is f which is worse than d)

- according to the fourth point of view, c is approved (and so is a which is better than c) whereas e is not (and neither is f which is worse than e)

The shift in the burden of proof allows the jury to support its claim (here, the result of the sorting procedure) with its chosen arguments. The length n of an explanation offers an indication of its cognitive complexity and the amount of information disclosed to the auditor. Therefore, we would instead provide the shortest possible explanations and strive to mention a few points of view as possible. Obviously, an explanation must reference a specific point of view at most once, so $n \leq |N|$. Unfortunately, we showed that one might require all points of view in a complete explanation, even in situations with relatively few alternatives.

Auditing conformity. We now wish to justify the committee's decision on a candidate $x \in \mathbb{X}$. As we have seen in the previous section, a complete explanation of the assignment of x implies disclosing much information related to the other candidates, which might not be acceptable. A possible solution is for a committee to base their decision on reference cases, an assignment $\alpha^* : \mathbb{X}^* \to \{$ GOOD , BAD $\}$, e.g. compiling past decisions that are representative of its functioning mode. In order to get rid of the influence of the other candidates, we are looking for *necessary assignments* given these reference cases.

Example 4.18.

We consider the alternatives a, b, c, d, e, f and their assignment $\alpha^\star$ have a reference status, and we are interested in deciding on the assignment of two candidates, x, y such that:

$$a \succ_1 f \succ_1 b \succ_1 e \succ_1 c \succ_1 y \succ_1 d \succ_1 x$$
$$e \succ_2 b \succ_2 y \succ_2 c \succ_2 d \succ_2 a \succ_2 f \succ_2 x$$
$$f \succ_3 a \succ_3 d \succ_3 b \succ_3 y \succ_3 x \succ_3 e \succ_3 c$$
$$d \succ_4 a \succ_4 c \succ_4 e \succ_4 x \succ_4 y \succ_4 f \succ_4 b$$
$$c \succ_5 y \succ_5 e \succ_5 b \succ_5 f \succ_5 x \succ_5 d \succ_5 a$$

It is not possible to represent the assignment $(x,$ GOOD $)$ together with the reference assignment α. Thus, x is necessarily assigned to BAD On the contrary,

> both assignments (y, GOOD) and (y, BAD) can be represented together with α.

Let us discuss in what follows the case of the necessary decision. We refer the reader to [Belahcene et al., 2018b] for the second case, where y is in an ambivalent situation.

An explanation of the *necessity* of an assignment is intrinsically more complex than that for its *possibility*: one needs to prove that it is not possible to separate all pairs of GOOD and BAD candidates on at least one point of view. The proof relies on some deadlock that needs to be shown. Formally, this situation manifests itself in the form of an unsatisfiable boolean formula. The unsatisfiability of the entire formula can be reduced to a $\subseteq$-minimal unsatisfiable subset of clauses (MUS), commonly used as certificates of infeasibility. It can also be leveraged to produce *explanations* (e.g. [Junker, 2004]). In the case of the necessary decisions by approval sorting with a reference assignment, any MUS pinpoints a set of pairs of alternatives in $(\alpha^{-1}(\text{GOOD}) \cup \{x\}) \times \alpha^{-1}(\text{BAD})$ that cannot be discriminated simultaneously according to the points of view.

Example 4.19.

Consider the subset of alternatives c, d, e, f, x, and assume x to be assigned to GOOD .

Each pair in $GB := \{(c, e), (x, d), (x, f)\}$ needs to be discriminated from at least one point of view in N, but this is not possible simultaneously: i) none of the pairs in GB can be discriminated neither from the first, the second nor the third point of view, as the overall GOOD alternative is deemed worse than the BAD one. ii) no more than one pair in GB can be discriminated according to each point of view among $\{4, 5\}$, and there are more pairs to discriminate than points of view.

The pattern of deadlock illustrated by Example 4.19 can be generalized and formalized into an argument scheme. Such an argument is a sufficient condition for the infeasibility of representing the given assignment in the non-compensatory model, which yields the *conclusion* that the candidate x is necessarily assigned to the other category.

To conclude, the proposed solutions stem from an original take of the dual notions of *possibility* and *necessity*, often used in so-called robust optimization, decision making [Greco et al., 2010] or voting contexts [Boutilier and Rosenschein, 2016] to account for incomplete information, conveying epistemic stances of skepticism or credulousness. Instead, we use them to describe the leeway left to the committee in setting its ex-

pectations: the decisions taken are bound from above by possibility, described as the feasibility of the Inv-NCS problem related to their decision, and from below by necessity, described as the infeasibility of the Inv-NCS problem simultaneously related to the reference cases and impossible assignments.

4.3 Summary

In this chapter, we presented our contributions to augment decision-aiding systems with explanation capabilities by using tailored "explanation schemes", i.e. argument schemes [Walton, 1996] dedicated to specific decision models to be used with explanation purpose in our context of decision-aiding. Just like argument schemes, explanation schemes can be seen as operators capturing prototypical reasoning patterns, i.e. a specific decision model in our case. In this context, one specific interest of these schemes is that, by splitting the reasoning process into smaller grains, they provide a natural building block (which the user can quickly grasp) for explanation lines. Moreover, providing an argument scheme along with the result (decision, recommendation) opens the possibility of discussing or challenging this result. This is made possible through what is called critical questions [Walton, 1996], a tool associated with argument schemes representing attacks or criticisms that, if not answered adequately, falsify the argument fitting the scheme (see Section 5.1). In our setting, the criticism may point out (implicitly or explicitly) elements perceived as missing or wrong in the reasoning steps. Indeed, the decision maker (DM) may challenge that a preference between two alternatives is not the right one. The consequence is that either it is possible to derive a new conclusion with this new information, or the DM's statements express conflicting preferences. Thus, the challenge of finding a principled way to deal with inconsistency in an accountable manner needs to be addressed (see Section 5.3). Smoothly interleaving explanation and recommendation calls for mixed-initiative systems (see Section 5.3), where the user may be active in challenging the system. Finally, the question of how the effectiveness of such systems should be evaluated (beyond their theoretical properties) remains largely open (see Chapter 5).

Interactive Recommendations and Explanations for Decision Support

5.1 Dialectical Tools for Decision Aiding

In the previous chapters, we presented our contributions for providing efficient and theoretically well-founded tools for both the preference elicitation task and explaining or justifying the outputs of the decision-aiding process. For recall, and as illustrated at the top of Figure 5.1, a decision-aiding process is an interaction between a human analyst (expert) and a human decision-maker, where the analyst aims to guide the decision-maker in building and understanding the recommendations of a particular decision problem.
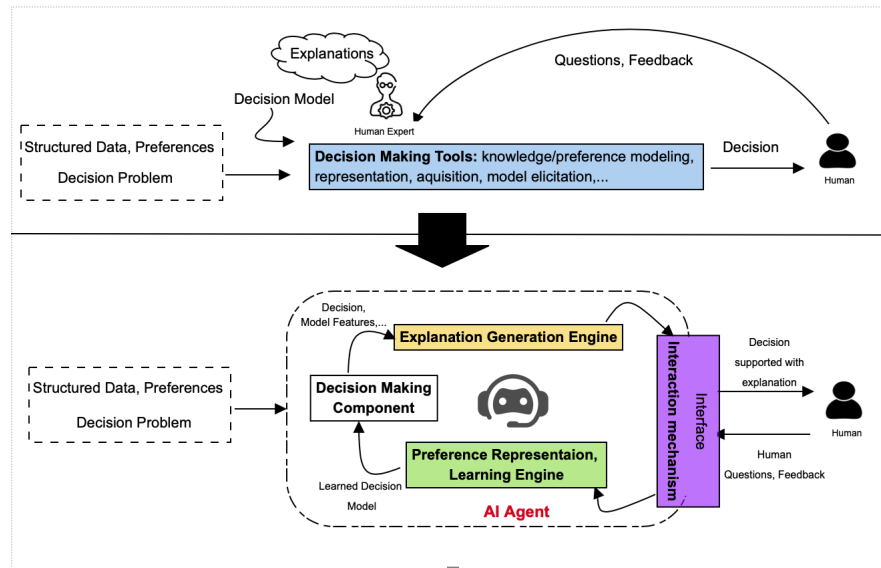

Figure 5.1: Dialectical vision for MCDA

Nowadays, decision-aiding situations are pervasive: they can occur in situations where the analyst's role is taken by a non-expert, even in some extreme cases by an artificial agent. This means that the artificial agent should ideally handle several

aspects – such as learning the preferences, structuring the interaction, providing an explanation, and handling the user feedback, . . . – usually delegated to the human analyst. Under such perspectives, our long-term research project is to design artificial agents, as illustrated in the lower half of the Figure 5.1 able to serve as analysts for various meaningful decision-aiding contexts. These agents will have different capacities (see red boxes in Figure 5.1). During the last years, we have focused our efforts on two components, elicitation and explanation engines, seeking to provide tools for each independently. The "Preference Learning Engine" has the task of setting up the model assumptions to work with for constructing the recommendation. It uses, for instance, the different algorithms proposed in Chapter 3 depending on the decision situation and the preference information (user profile). As we shall see later in this chapter, introducing explanation capabilities and interactive features with a human user will raise new issues in designing efficient tools for preference elicitation. On the other hand, the "Explanation Generation Engine" aims to provide the justification (or explanation) given to the user on the proposed items or facts inferred by the agent during the interaction. We can rely, for instance, on the different proposals described in Chapter 4. Finally, even if the Figure 5.1 was conceived with the multi-criteria decision aiding framework vision, we do not doubt that it can be adapted to any setting where the notion of preferences (human user) is at stake. Some ideas are discussed in the rest of the chapter.

Therefore, if we are to automate (some part) of the process, it is essential to understand more clearly how the tasks handled by a human analyst can be integrated into a tool. More precisely, it would be helpful to design and implement formal tools to support this interaction between the artificial agent and the human user. Our target is to answer the following question:

> *How to equip an artificial agent with adaptive behavior and model the system's reasoning to allow "efficient" interaction with a user within a decision-aiding situation?*

Although we have focused most of our work on explainability and preference elicitation, we have conducted the first reflection on the question of designing this interaction between an artificial agent and a human user (the box "interaction mechanism" in Figure 5.1). We grounded on dialectic models from the multi-agent systems field, specifically argumentation-based dialogues [Walton and Krabbe, 1995; Black et al., 2021]. Our different proposals, summarized in Table 5.1, have been carried out mainly during our PhD thesis [Ouerdane, 2009] and we intend to continue and extend it in the coming years. A promising continuation is the one started in the PhD of Amoussou [(in progress)].

Dialectical interaction models have gained tremendous popularity in recent years in the multi-agent community. Many protocols have been put forward to tackle

Approach	References
Argumentation-based interaction	[Ouerdane et al., 2008] [Ouerdane, 2009] [Ouerdane et al., 2010] [Ouerdane et al., 2011] [Labreuche et al., 2015]

Table 5.1: Our contributions to adaptive interaction

different types of interaction [Walton and Krabbe, 1995]. It is clear that these protocols offer greater expressivity than simple feedback (since recommendations can be challenged and justified). Our work follows this trend of research and studies a type of interaction whose specifities have seldom been studied. More precisely, we investigated relying on argumentation-based dialogue to formalize the interaction between a decision-maker and an artificial analyst within a decision-aiding process. Argumentation theory is a rich, interdisciplinary area of research across philosophy, communication studies, linguistics and psychology. Its techniques and results have found a wide range of applications in both theoretical and practical branches of AI and computer science [Bench-Capon and Dunne, 2007; Simari and Rahwan, 2009].

In recent years, argumentation theory has gained increasing interest in the multi-agent systems (MAS) research community. It can be used: (i) to specify autonomous agent reasoning (belief revision, decision making under uncertainty, ...): it provides a systematic means for resolving conflicts among different arguments and arriving at consistent, well-supported standpoints; and (ii) as a vehicle for facilitating agent's interaction. It naturally provides tools for designing, implementing and analyzing sophisticated forms of interaction among rational agents [Amgoud et al., 2000; Atkinson et al., 2005; Charif-Djebbar and Sabouret, 2006; Black et al., 2021]. More recently, argumentation theory has received particular attention in the XAI field (see [Čyras et al., 2021; Vassiliades et al., 2021]) as it naturally provides a means to construct explanations and justifications.

While the link between decision-making and argumentation has been investigated over several years [Atkinson et al., 2006; Amgoud and Prade, 2009; Fox and Parsons, 1998; Kakas and Moraitis, 2003; Müller and Hunter, 2012], the decision-aiding setting itself has been little studied. Fore recall, a decision aiding context implies the existence of at least two distinct actors (the user and the expert) both playing different roles; at least two objects, the user's concern and the expert's (economic, scientific or other) interest to contribute; and a set of resources including the user's domain knowledge, the expert's methodological knowledge, money, time... The ultimate objective of this process is to come up with a consensus between the user and the expert [Tsoukiàs, 2008]. For implementing and formalizing this dialogue, we have put in place several tools to: i) conduct the interaction, ii) manage the various preference models, and iii)

allow critics and feedback from the user. These different aspects are discussed in what follows.

5.1.1   Conducting the interaction though a dialogue game.

A first step towards formalizing such a discussion is our work [Labreuche et al., 2015], where a dialogue game is proposed to formalize the interaction representing a decision-aiding situation, involving the exchange of different types of preferential information, as well as other locutions such as justification. We have two players: the DA (Decision Aider: the artificial agent) has the aim of constructing a solution to a given decision problem. The DM (decision-maker: the human user) expresses his preferences through feedback and has to be convinced by the solution. Moreover, during the dialogue, the DA constructs a Knowledge Base (KB) composed of the Preference Information (PI) provided by the DM and the accepted statements. The protocol for our dialogue model is depicted in Figure 5.2, where grey nodes are for the DM, white nodes for the DA.

$$\phi_3 \subseteq \phi_1,\ \phi_3 \subseteq \phi_2$$
$$\phi_4 \subseteq \phi_1,\ \phi_4 \subseteq \phi_2,\ \phi_4 \subseteq \phi_5$$
$$\phi_5 \subseteq \phi_3$$
$$\phi_6 \subseteq \phi_4$$
$$\phi_7 \subseteq \phi_6,\ \phi_7 \subseteq \phi_5 \text{ or } \phi_7 \text{ conflicting with } \phi_5$$
$$\phi_8 \subseteq \phi_1,\ \phi_8 \subseteq \phi_2,\ \phi_8 \subseteq \phi_5$$
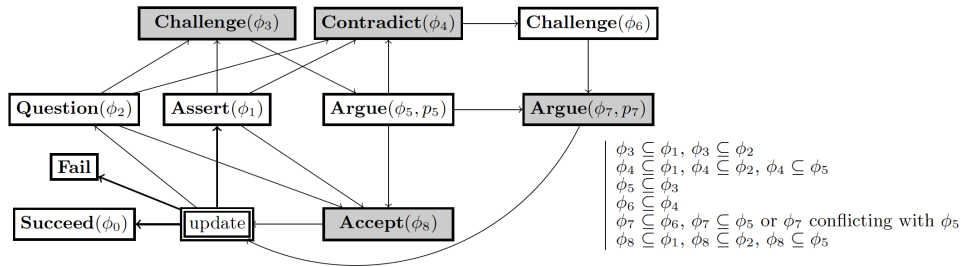
Figure 5.2: Successive speech acts at each iteration

Briefly, each node in this graph is a locution, except for "Update". This latter enables the DA to analyze the exchanges made during the last iteration of the dialogue, update the KB and construct the proposal for the next iteration. The outgoing arcs from a node indicate the possible following locutions. A dialogue under this protocol is composed of several iterations. Each iteration starts from the node "update" and is organized around an assert(ion) or a question made by the DA and the feedback of the DM. Among the results, we prove that this protocol satisfies desired properties, in particular termination and efficiency (in the sense that the recommended option is indeed among the most preferred of the decision-maker).

In this work, we mainly focus on constructing an interaction protocol that specifies the rules and conditions under which we can have a "coherent" interaction in a decision-aiding context where the initiative is sometimes left to the user (e.g., ask for an explanation). Different perspectives are possible besides the assumptions assumed

to construct this first proposition that can be relaxed. The first one concerns the preference elicitation process. Indeed, we use default weights and scores to handle incomplete preference statements instead of relying on a specific technique/algorithm of elicitation. Thus, it would be interesting to design a protocol that will consider the elicitation task and generate recommendations supported by explanations. As we shall see in Section 5.3 interleaving elicitation and explanation raises new questions. Another interesting perspective is to go through the implementation of such a protocol and conduct experiments to validate the approach (see Section 5.3). A further challenge is exploring how the user's preference information will be captured and integrated into the system. Of course, how to present the recommendation and the supporting explanation is an interesting issue, too (see Section 5.2 for a discussion). Finally, as we shall see at the end of this document, this question of designing dialogues for an artificial agent within an XAI context is also challenging for other application domains.

5.1.2 Managing various preference models.

In classic decision theory works, and given a decision situation, a decision analyst first chooses the model based on the desired properties (axioms satisfied by the model) and then proceeds to elicitation. This task will aim to set up the model assumptions to work with for constructing the recommendation. However, in a practical context, such a preliminary assessment might not be feasible. Thus, rather than making an assumption that may later be found to be incorrect (as an example: the weighted mean model is often used in many systems but without an explicit justification), our idea is to simultaneously reason with several possible models and let the system decide the one appropriate to the current user.

More precisely, we proposed in [Ouerdane et al., 2010; Labreuche et al., 2015] an approach that allows the artificial agent to use a variety of decision models (able to encompass most decision situations) to build its recommendation (as opposed to adjusting the parameters of a single model). To account for this, an axiomatic approach is adopted, where the use of a model is triggered by a set of properties that should the decision maker's preferences be fulfilled. In other words, to adapt to different DMs, the DA will use a range of decision models Π, where a set of properties identifies each model. Such properties correspond to some characteristics of the DM's preferences, corresponding to a set of conditions supporting the use of a given model.

For illustration, let us consider the following family Π of models: Simple Majority model (noted π_{SM}), Simple Weighted Majority model (π_{SWM}), Mean model (π_M) and Weighted Sum model (π_{WS}). Therefore, we denote by Q the set of properties. For a given model $\pi \in \Pi$, each property can be either satisfied or not. For illustration, we will consider the set of properties Q that include: (1) Cardinality of the model (*car*): it means that the specific difference in performance values makes sense (when this property

is not satisfied, only the ordering of options is relevant for comparison). (2) Non-Anonymity of the model (nan): it suggests that criteria are not exchangeable (when this property is not satisfied, all criteria are exchangeable). With $Q = \{car, nan\}$, we can describe the four decision models $\pi_{SM}, \pi_{SWM}, \pi_M, \pi_{WS}$. On top of the two properties, Cardinality (car) and Non-Anonymity (nan), let us introduce a *veto* property (vet) saying that there is a veto criterion. One can readily see that not all combinations of properties yield a relevant decision model. Figure 5.3 shows the set of relevant properties. For instance, the "outranking model" (noted π_{OR}) corresponds to property vector $(\bot, \top, \top)$: it is ordinal but uses criteria weights and veto criteria. On the other hand, property vector $(\bot, \bot, \top)$ has no relevant corresponding model as it satisfies only veto. A similar situation arises for $(\top, \bot, \top)$ and $(\top, \top, \top)$ as a cardinal model (weighted sum) able to represent a veto criterion subsumes to a dictatorial rule (only one criterion counts), which is not very interesting and can be represented by π_{OR}.

$$\pi_{OR} \; (\bot, \top, \top) \qquad\qquad \pi_{WS} \; (\top, \top, \bot)$$

$$\pi_{SWM} \; (\bot, \top, \bot) \qquad\qquad \pi_M \; (\top, \bot, \bot)$$

$$\pi_{SM} \; (\bot, \bot, \bot)$$
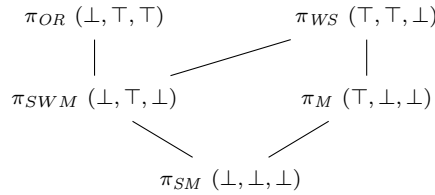
Figure 5.3: Structure Q with three properties

The set Q is used to guide the navigation among the different models (or associated subsets of properties), depending on the properties that are currently satisfied or contradicted.

Let us consider for illustration an excerpt of an exchange between a DA and a DM as depicted in Example 5.1 (see Chapter 1). This exchange has as input the comparison of the options over each criterion provided by the DM.

Example 5.1

Let us consider the following situation for illustration. Suppose that a decision-maker specifies that he has to rank four options $\{a, b, c, d\}$ (say, bikes to be deployed for sharing in a big city). Each bike is evaluated on the set $\{c_1, c_2, c_3, c_4, c_5\}$ of criteria (say, price, weight, aesthetic, gears, dimension). The comparison of the options over each criteria (where $x \succ_{c_i} y$ means that option x is strictly preferred to y on criterion c_i) is as follows:

$$c_1: d \succ_{c_1} a \succ_{c_1} c \succ_{c_1} b;$$
$$c_2: d \succ_{c_2} a \succ_{c_2} b \succ_{c_2} c;$$
$$c_3: b \succ_{c_3} c \succ_{c_3} a \succ_{c_3} d;$$
$$c_4: c \succ_{c_4} b \succ_{c_4} a \succ_{c_4} d;$$
$$c_5: b \succ_{c_5} a \succ_{c_5} c \succ_{c_5} d.$$

(1) DA: I recommend that $b \succ a \succ c \succ d$.
(2) DM: Why $b \succ a$?
(3) DA: b is better on a majority of criteria (c_3, c_4, c_5).
(4) DM: I see, but still I would prefer a to b
(5) DA: Why?
(6) DM: Because a is better on the price and weight (c_1, c_2), these are very important.
(7) DA: Fine. I still recommend b over c.
(8) ...

At the first iteration (1), the DA generates a first recommendation from the partial preferences of the DM and provides a justification at iteration (3). In this iteration (3), solely based on comparisons provided by the DM and without any other information (i.e. we do not proceed to the elicitation of more information), the DA assumes that the model is π_{SM} (in the Figure 5.3 node $(\bot, \bot, \bot)$). Note that the agent made this assumption to start the interaction. The idea, as discussed previously, is that during the dialogue, if we get a piece of additional information and this information contradicts the assumption, we update the decision model. This is the case at iteration (7), where the model π_{SWM} is used due to statements $[c_1 = strong]$, $[c_2 = strong]$. Technically, we move in the Figure 5.3 from node $(\bot, \bot, \bot)$ to the node $(\bot, \top, \bot)$ on the basis that c_1 and c_2 are more important than the other criteria, and thus the Non-Anonymity (nan) property should be taken into account. Note that the inference of the comparison among options is consistently constructed even though the model is changing, thanks to the relation between the models and the related properties.

To navigate among the different nodes based on the responses of the decision-maker during the interaction, we established a list of "critical responses (questions)" borrowed from arguments schemes [Walton, 1996] (see the following section). Such responses offer a way to identify what property is challenged or which should be taken into account.

5.1.3 Allowing critics/feedback through Critical Questions.

During the interaction with the system, it is necessary to provide the decision-maker means to communicate with the system and express his doubts about the conclusions and explanations (arguments) presented. Thus, the decision-maker is involved in developing the recommendation by pointing out those elements that appear missing or wrong in the reasoning steps assumed by the system. To this end, we borrowed a tool from argumentation theory named "critical questions". Indeed, our first objective by relying on argument (explanation) schemes is a knowledge representation exercise. By casting the reasoning steps under the form of argument schemes, we make explicit assumptions usually hidden for the decision-maker, hence allowing meaningful explanations. The second shows that argumentation tools

facilitate the revision/update occurring during such a process. Indeed arguments schemes come along with what we call *critical questions*. They represent attacks, challenges or criticisms that, if not answered adequately, falsify the argument fitting the scheme. Asking such questions throws doubt on the structural link between the premises and the conclusion. They can be applied when a user is confronted with the problem of replying to that argument or evaluating it and whether to accept it.

A first attempt to define what critical questions (responses) could be in a decision-aiding situation is our thesis work [Ouerdane et al., 2010, 2011, 2008]. For illustration, if we go back to our Example 5.1, at the turn (7), the DA generates a recommendation based on the reaction of the DM at turn (6), which through its response implicitly modifies the decision model under use. Indeed, the DM's response puts forward that Non-Anonymity (nan) property is no longer fulfilled, as he considers precisely two criteria (very important) in comparing a and b. We have identified the following set of possible responses that could lead to the assumption that the nan property should be taken into account:

- the criterion c_i is more important than the criterion c_j

- option x is better than option y on the coalition of criteria $\{c_i, c_j\}$

- if option x is preferred to y on the criterion c_i, it should be the same on the criterion c_j

- x is too bad (or better than anyone else) on the criterion c_j

In the Ex.5.1 the turn (6) is assumed to correspond to the second type of response.

Such responses were constructed by respecting the theory and concepts of decision-aiding methodology. However, we believe that an experimental study aiming at analyzing the decision-maker's behavior in a situation of decision support would probably confirm such responses and allow us to identify other more realistic and practical reactions. Such a study could also validate the properties specified in [Labreuche et al., 2015] and identify other natural features of the decision-maker preferences that we have not thought about. Moreover, the use of critical questions is not restricted to challenging the preference aggregation procedure but is a promising tool to elicit preferences (see Section 5.3).

5.1.4 Next steps

To summarize, the construction of the different components (see Figure 5.1) of the artificial agent depends on the decision situation faced by the user. Such a situation will clearly impose a particular decision model in the classical setting. However, our

idea is that rather than making an assumption that may later be found to be incorrect (as an example: the weighted mean model is often used in many systems but without an explicit justification), we suggest simultaneously reasoning with several possible models and let the system decide the one appropriate to the current user. Therefore, it is clear that elicitation/explanation/interaction (dialogue) algorithms should be adapted to the considered situation.

A first baseline version of our artificial agent can be the one with: explanation patterns [Belahcene et al., 2017b] and an elicitation mechanism [Viappiani and Boutilier, 2009] for the additive utility model, with the interaction model of [Labreuche et al., 2015], where the aim at the end is to articulate these components to provide an integrated model. This baseline is still ongoing work, as the integration is not an easy task, but we hope we can get the first version within Amoussou [(in progress)]'s PhD .

Finally, beyond this basic version, putting together the different pieces to build this artificial agent for decision support opens up new work areas with new opportunities for collaboration with new colleagues. These perspectives are discussed in the following. We want to draw attention to the fact that the rest of the document is not intended to have an exhaustive state of the art or to detail the contributions, but to give the few avenues on which we wish to work in the coming years.

5.2 Explanation Schemes: Generation and Evaluation

In our different proposals for providing explanations to justify recommendations (see Chapter 4), we have concentrated our efforts essentially on two MCDA models: the additive model and the NCS model. Moreover, neither natural language generation nor in vivo experimentation were investigated in the different contributions. For instance, the complexity of explanations was assessed through proxies, such as length or number of premises. Several perspectives can be envisaged to enrich our work in this perspective of equipping an artificial agent with explanatory capacity.

5.2.1 New explanation schemes/patterns

In MCDA, various unexplored models remain for which the questions of constructing explanation schemes are relevant. We aim to enrich our catalog with other explanation (argument) schemes by considering additional decision models and situations. Such a catalog will offer the artificial agent the ability to construct the appropriate explanation according to the decision situation and thus a decision model. Moreover, even if our research work has long focused on models or methods from the field of multi-criteria decision aiding, our ambition is to open to methods and models in other areas such as Operation Research (OR) and Machine Learning (ML).

Explaining outputs of Optimization Systems. In this direction, we have already started within the PhD of Lerouge [(in progress)] a work in the OR field. In collaboration with Vincent Mousseau (MICS, CentraleSupélec), Celine Giquel (LISN, Université Paris-Saclay) and Decision Brain[1], we investigate the question of explaining solutions stemming from the Workforce Scheduling and Routing Problem (WSRP), an optimization problem, to an end-user. In brief, a WSRP can be described as follows: given a set of n mobile employees and a set of m geographically dispersed tasks, the problem consists in building pairs of paths and schedules and in assigning a path-schedule couple to each employee defining which tasks he should perform, in what order and at what times. The objective is to design a family of path-schedule couples of minimum cost, which accommodates as many tasks as possible while satisfying a set of constraints [Castillo-Salazar et al., 2016]. For our purpose of explainability, the first proposition was to consider an instance of WSRP and a solution and allow the user to query the solution's relevance. With the help of our industrial partner, Decision Brain, we identified a bunch of questions that an end-user may ask. These questions are local - they relate to a part of the solution - and contrastive [Lipton, 1990]. This reduces the size of the calculation determining the explanatory content and *in fine* provides an explanation to the user in real-time. More precisely, we use polynomial algorithms using tools from local search or integer linear programming applied to small problems to compute an explanation. Finally, to be intelligible to the user, the explanation takes the form of concise text, written in a high-level vocabulary, and graphics (e.g. representations of the solution, performance indicators of the solution). This is ongoing work, and we aim to pursue it on different tracks. For instance, as we are dealing with a real-world case study with an industrial partner, it would be interesting to tackle the evaluation question. The idea is to conduct experiments with end users to get feedback on the relevance of the produced explanations. This raises different questions as discussed in Section 5.2.3.

Explaining outputs of ML models Regarding the ML direction, our first tentative on this subject will be carried out in collaboration with Hopia[2], Gianluca Quercini (LISN, CentraleSupélec), Myriam Tami (MICS, CentraleSupélec) and Paul-Henry Cournède (MICS, CentraleSupélec). Hopia is a start-up that offers a planning solution for healthcare institutions. Among the question that Hopia should consider to setting up optimized planning is to be able to establish the patient flows in a hospital system. To this end, the project aims to investigate data-driven methodologies that can assist in predicting/analyzing periodic behavior. More precisely, the ambition is to develop predictive models based on integrating several data on the patient and

---

[1]A French company specializing in optimization software development has several client companies who daily need to solve instances of WSRP https://decisionbrain.com

[2]https://hopia.eu

the hospital department and considering patient flows between departments. In addition to predictions, the model will need to incorporate a measure of uncertainty in the predictions (confidence intervals on the prediction) and accommodate incomplete data. In this context, different machine learning models will be considered. Therefore, to respond to the problem of trustworthy AI generated by using ML models and in a sensitive context such as health, the project will design tools for the interpretability and explainability of results appropriate to the context. In this perspective, we envisage adopting an interactive approach where the explanation will be a source of interaction to allow feedback, corrections and new information from the user (medical staff in this situation), thus enriching the learning phase. Indeed, as pointed out by [Lindsell et al., 2020] the successful use of AI tools in the health field depends not only on the progress of AI algorithms but also on the human in the loop which involves all stakeholders. This project is already initiated by a six months Master Internship at MICS started on 2 May 2022, on the subject "AI for predicting Patients Flow" funded by DataIA[3], under our supervision. In the following steps, it is envisaged to construct with Hopia a PhD subject and look for funding and a PhD Candidate.

5.2.2 Expressing and presenting an explanation?

In this context of generating explanations, another interesting and challenging question is *how to present (communicate) explanations to a user?* We believe that a promising direction is to approach the problem of explanation generation as a problem of planning [Cawsey, 1993], where the idea is to find the path that leads to the conclusion. Since our results identified several basic "operators" (under the form of argument schemes), it is thus tempting to adopt this stance and design an explanation planner for our decision-aiding setting. Several alternative plans with different explanation strategies can be represented, which may be triggered depending on the context and user feedback. This is planning under uncertainty since different user reactions may affect execution. The user may thus interrupt a line of explanation, for instance, because he cannot grasp a specific elementary step of the explanation, forcing him to backtrack to an alternative -hopefully better suited- one. This unified framework could pave the way for a potentially powerful mixture of approaches (using different types of argument schemes within the same line of explanation).

Moreover, we did not rely on Natural Language Generation (NLG) tools to express explanations for our different contributions. We aim to do so. Using the NLG tools will imply tackling all the aspects of the generation process in a principled way, from selecting and organizing the content of the explanation to expressing the chosen content in natural language. Text generation involves two fundamental tasks: a process that selects and organizes the content of the text (deep generation) and a process

---

[3]https://www.dataia.eu

that expresses the selected content in natural language (surface generation) [Reiter and Dale, 2000]. The challenge is to develop a complete computational model for generating explanation schemes tailored to the user's preferences.

Moreover, for the surface generation, the literature [Forrest et al., 2018; Alonso and Bugarın, 2019; Pierrard et al., 2019; Baaj and Poli, 2019] use mostly surface realizers like SimpleNLG [Gatt and Reiter, 2009] to produce textual explanations, despite some drawbacks. For instance, the latter does not easily handle the inclusion of notions or concepts expressing uncertainty, probabilities or confidence in the text. On the other hand, the NLG is a separate domain that is not necessarily mastered by the people who implement XAI systems, which explains why the link between the two is still difficult to establish, especially when it comes to extracting the relevant information from the underlying model. We believe that there is a need to build a bridge between the extraction of the content of the explanation and the construction of the textual representation.

To meet this need, we have the idea to design a *semantic representation* of the content of the explanation [Baaj et al., 2019]. Indeed, from our point of view, the explanation generation process can be viewed as a sequence of three main tasks, namely: (i) content extraction from an instantiated AI model, (ii) semantic representation of this content and finally, (iii) text generation using NLG techniques [Baaj et al., 2019]. More precisely, content extraction is specific to each AI model (neural networks, expert systems, etc.): it takes as input the instantiated model, i.e. all the values of the model for given inputs (e.g. the values of the weights for a neural network, the execution trace for an expert system, etc.). Conversely, the other components are common to all models so that the mechanisms can be mutualized. This decomposition of tasks can also help the evaluation by allowing, for example, to evaluate the content of the explanation without considering the text generation. The ambition is to build a semantic representation independent of the AI model. Thus, any specialist of an XAI model will be able to represent his explanation without worrying about the textual part. This perspective is joint work with Jean-Philippe Poli (CEA List), where our ambition is to propose a formal structure that explicitly links the concepts (components) of the explanation to each other and allows the representation of logical and causal relations between these elements. This requirement has been emphasized by [Chari et al., 2020], where it is claimed that such a representation can contribute to a better understanding of explanations and be beneficial for constructing AI systems that will help users through a so-called "distributed cognition" approach [Hollan et al., 2000]. The system generates explanations aligned with the users' needs in this context. The first tentative in this perspective was addressed in [Baaj, 2022], but there is still work to develop a convincing proposal.

5.2.3 Evaluating and Assessing explanations

When dealing with systems that emphasize explainability, it is essential to assess how pertinent explanations are correct. Until now, in our different contributions, the complexity of explanations was evaluated through proxies, such as the length or the number of premises.

Different works in psychology have discussed how a human user could evaluate or perceive an explanation. For instance, [Miller, 2019] reviewed the main factors that play a role in the human assessment of a "good" explanation. The authors state that a good explanation needs to be *coherent*. That means that it must be consistent with the end-users knowledge [Thagard, 1989]. In [Hoffman et al., 2018] different methods for evaluating (1) the goodness of explanations, (2) whether users are satisfied by explanations, (3) how well users understand the AI systems, (4) how curiosity motivates the search for explanations, (5) whether the user's trust and reliance on the AI are appropriate, and finally, (6) how the human-XAI work system performs, are discussed. On the other hand, Read and Marcus-Newhall [1993] consider that users prefer *simpler* explanations (those that cite fewer causes) and more *general* explanations (those that explain more events). Also, people do not usually judge an explanation based on its probability but rather on its usefulness and relevance [McClure, 2002].

Several solutions have been proposed in the XAI literature to assess or evaluate explanations [Mohseni et al., 2021]. The authors classify them into three methods: (i) Application-grounded evaluation, where an expert directly evaluates how good an explanation is, and (ii) Human-grounded evaluation, a human is asked to perform simple experiments that are still linked to the target. For example, one or several humans could be asked to select the best explanation among several of them, and (iii) Functionally-grounded evaluation, where the idea is to assess the explanations of one model with another model that has been previously validated as an explainable model. Following the human-grounded evaluation, we have initiated a first work with Jean-Philippe Poli (CEA List). This work focused on the generation and the evaluation of the explanation [Poli et al., 2021]. In this proposal, an explanation is a sentence in natural language dedicated to human users to provide clues about the process that leads to the decision: the assignment of the label to image parts. We focus on semantic image annotation with fuzzy logic that has proven to be a helpful framework that captures both image segmentation imprecision and the vagueness of human spatial knowledge and vocabulary. In this work, we presented two algorithms for textual explanation generation of the semantic annotation of image regions. To compare the two approaches, we evaluated both of them. In this aim, we use the questionnaire presented in [Baaj and Poli, 2019]: it is based on 17 questions organized into three categories: natural language, human-computer interaction and content and form. Each question is evaluated with a Likert scale (from 1 "strongly disagree" to 5 "strongly

agree"). Our panel consists of 40 respondents, with 20 medical staff members (medical doctors, surgeons, nurses, radiologists), the other half being computer scientists (6) and other various non-medical professionals (14). Among the results, the *order* of the items inside an explanation seems to be essential for the end-users. *conciseness* is a criterion of paramount importance.

Clearly, work still needs to be done to implement the most acceptable way to evaluate our several explanation schemes. We will take advantage of our previous work and from both psychology and XAI literature to set up experimental protocols and define criteria that seem relevant regarding the decision-aiding situation. The goal will be to validate the relevance of our explanation schemes from the point of view of a human user.

5.3   Interactive explanation and inconsistency management

While the classical incremental elicitation methods already involve an interactive process whereby the system asks queries to the user (see for instance, [Benabbou et al., 2017; Gilbert et al., 2017; Perny et al., 2016; Adam and Destercke, 2021]), there are new challenges when one wants to integrate explanation facilities.

5.3.1   Mixed-initiative interaction

The current systems equipped with explanation features typically produce justification at the very end of the process– together with their final recommendation. We believe that an adequate explanation cannot be one shot and involves an iterative communication process between humans and artificial agents. As humans can easily be overwhelmed with too many or too detailed explanations, the interactive communication process helps understand the user and identify user-specific content for the explanation. Moreover, cognitive studies [Miller, 2019] have shown that an explanation can only be optimal if it is generated by considering the user's perception and belief.

Under such a perspective, we think that a mixed-initiative system [Horvitz, 2000] where elicitation, recommendation and explanation are tightly interleaved, is required. According to [Horvitz, 2000], mixed-initiative systems refer "broadly to methods that explicitly support an efficient, natural interleaving of contributions by users and automated services aimed at converging on solutions to problems". The management in such systems is non-trivial, as it must be possible to decide which side should be granted the initiative during the interaction. This implies carefully designing a protocol which decides exactly how and when the initiative should be given to the user or kept by the system and how the different commitments can be agreed upon or challenged.

In our context, one key issue will be identifying when exactly explanations can be triggered by the system or asked for by the user. A further difficulty is that the nature

of explanation patterns may vary. Some explanations will require a specific interaction with the user, others will be planned beforehand, and visual explanation may be part of the process. A careful analysis of the proposed protocols will guarantee termination or efficiency properties of the protocol under natural assumptions of the user's behavior. Unfortunately, often the user cannot be assumed to respond consistently throughout the interaction, which leads us to integrate means to manage inconsistency (see the next point).

Moreover, as discussed in the previous section, an interesting tool for interaction and getting feedback and new information from the user is the critical questions attached to an argument scheme. In Chapter 4 we established various argument schemes to support different types of recommendations (assignments, choices, pairwise comparisons); we plan to rely on critical questions to evaluate such schemes. This perspective can keep the user in the loop, which is often essential in a decision situation. Moreover, a thorough study should be done, theoretically and by experiment, to see to what extent such a tool could benefit the preference elicitation process.

5.3.2   Modeling and managing inconsistency

To produce a recommendation, the system questions the user to elicit her preferences and fit them into a model. Based on these preferences, the system can produce a recommendation. However, because the recommendation itself can be very large (think of a ranking involving all the options), it is useful to allow incremental partial and/or factored recommendations to be made throughout the interaction, on which the system will seek the agreement of the user (e.g. "do we agree that product p is better than any product which color is red?", or " do we agree that subset of options p_1, p_2, p_3 should not be considered as the product of choice?"). When the system puts it forward, the user can critique it (preferences may be adjusted, corrected, the option may not be feasible, or not available anymore, etc.) or asks for a justification, which the system must provide. As a result, the system must deal with the inherent *revision problem* induced by the possibly incoherent statements (either among themselves or with the user assumed preference model).

More precisely, such "inconsistencies" may occur when, for instance: the DM's statements express conflicting preferences, the DM's point of view is evolving during the interaction process, and the DM's reasoning is incompatible with the principles and properties underlying the preference model, etc. Therefore, we aim to investigate modeling and handling inconsistency during an interaction between an artificial system with a user. Different issues arise: How should the system behave in the presence of inconsistency in the situation where a (family of) model(s) cannot restore the DM's preferences? Should we revise the expressed preferences? Should we change the model? Thus, on what principles? How to conduct the elicitation process by taking into account the in-

consistency? Actually, on the one hand, neither active learning nor complete elicitation strategies deal with the question of revising the model. On the other hand, generating an explanation adds complexity to this question as it becomes legitimate to seek to find/keep the information that will allow the construction of "good" explanations at the end. We could rely on different strategies.

- Constructing maximally consistent subsets of statements. For instance, an approach that identifies minimal inconsistent sets of preference statements was proposed by [Mousseau et al., 2003], i.e., subsets of statements that, when removed, lead to a consistent system. Identifying such subsets would indicate the reason for the conflicting information. In the same spirit, we can think of using logical formulation and try to identify, for instance, a minimal unsatisfiable subset of clauses (MUS) [Junker, 2004].

- Relying on a numerical estimation of inconsistency, such as a belief function. Destercke [2018] has proposed a general setting based on evidence theory allowing to deal with inconsistency and uncertainty in user feedback, which seems attractive from the perspective of revising a model. With this perspective, it will be an opportunity to collaborate with Sébastien Destercke (Heudiasyc, Université de technologie de Compiègne, CNRS).

- Relaxing the aggregation model. One way to interpret the inconsistency is that the actual decision model cannot represent the user's preferences. We have proposed a first solution based on an axiomatic approach toward relaxing/changing the decision model. We envisage continuing to investigate this issue in the future. In addition to the axiomatic approach, we may consider an automatic incremental model selection: this is a challenging approach, as the learning process of the model is intertwined with that of learning the preferences.

- Relying on explanatory dialogue. Finally, an interesting direction to solve inconsistency could be the approach described in [Arioua et al., 2016, 2017], where the authors propose a framework of inconsistency handling through knowledge acquisition through an explanatory dialogue. More precisely, by relying on argumentation-based dialogue. The approach is based on interacting with a user to acquire new knowledge and feedback to remove inconsistencies. This avenue aligns with our vision of using argumentation and explanation through dialogue. Thus it could be attractive to see to what extent it could be applied/extended to our setting.

5.3.3　New perspectives for preference learning and elicitation

The preference elicitation task aims to correctly represent the user's preferences through a given model to fit the user's rationality. As was pointed out by (Boutillier,

2013): "no decision support system can recommend decisions without some idea of what are the preferences of the user. This information cannot be coded into the system in advance and raise the preference bottleneck: how do we get the preferences of the user *into* the decision support system?"

Our ambition is to endow the virtual agent with tools to capture incrementally the user's preferences and feedback (contradicting a previous assertion, asking for an explanation, etc.) while minimizing at the same time the cognitive effort of the user. Under these perspectives, a challenging issue is a computational aspect. In particular, we want to provide elicitation techniques that can cope with inconsistent or "noisy" user feedback by automatically adjusting the model to the preference information provided by the user.

We have already started work in this direction concerning the computational aspect by proposing new tools based on logical formulations that have shown superior performance to those of mathematical programming, a classical formalism in decision theory. We intend to continue in this direction for other models of multi-criteria decision aiding. In addition, in the midterm, we would like to investigate if it is possible to build tools that combine the interpretability of MCDA models and the efficiency of machine learning algorithms. A trend in AI is the hybridization of the so-called symbolic mechanisms and those of ML. It will be interesting to see how this hybridization can be designed in a multi-criteria decision-aiding setting and which mechanisms we can implement. This perspective will be the occasion to collaborate with some colleagues in ML in the lab. Concerning the inconsistency part, several tracks were evoked in the previous paragraph. Investigating how to efficiently couple these tools and the elicitation algorithms will be a question.

5.3.4   Interaction: validation and evaluation

Designing an artificial agent with explanation features for decision-aiding purposes will require a validation phase. In other terms, how to experiment and/or practice a decision-aiding situation with the help of an artificial agent endowed with an explanatory capacity. Thus, we need to carefully elaborate: (i) what can be "good" indicators or criteria to assess and validate the results. For instance, one can intuitively assess the interaction's convergence by making a compromise between accepting (or not) a recommendation and the time spent to obtain the agreement. However, it is less clear how to assess the impact of introducing an explanation within a recommendation). Moreover, (ii) a methodology or a framework of how validation should be implemented. In other terms, how to experiment and/or practice a decision-aiding situation with the help of an artificial agent endowed with an explanatory capacity.

5.4   Towards Decision Aiding for Collective Decision

We have always dealt with decision situations with the hypothesis of a single decision-maker (end-user). We still have several interesting and rich avenues to explore with many collaborations in prospect. Besides, in the longer term, we would like to extend our work to the multi-decision maker, the multi-participant context. An exception is our work in [Belahcene et al., 2018b]. In this paper, we were interested in the problem of accountability of decisions issued from a non-compensatory sorting model (NCS) [Bouyssou and Marchant, 2007a]. Two situations have been mainly studied. In the first one, a committee must justify its decision as a possible NCS assignment. The second situation arises when the assignment of a new candidate is necessarily derived from jurisprudence. In this work, even we have a committee (a group), but the explanation issue has been treated to account for the committee's decision-making process towards an external entity. Therefore, we wish to deal with the situation where the decision concerns a group of individuals, and thus we need, for instance, to explain that the solution found is fair for the whole group.

In a collaborative decision problem, one seeks to aggregate different participants/agents' preferences on given alternatives to reach a joint decision. Examples of such problems include voting problems such as the election of political representatives or the choice of projects to be funded in a municipality, resource allocation and fair sharing problems such as the assignment of papers to reviewers in a conference or the assignment of students to courses, or coalition-building issues such as the assignment of undergraduates to higher education institutions or the formation of student groups for projects. The study of collective decision-making falls within the computational social choice [Brandt et al., 2016], a sub-field of artificial intelligence that aims to analyze collective decision-making from an axiomatic and algorithmic perspective. In this context, participants can exchange information, oppose other participants, ask for clarifications/justifications, revise their views, establish strategies, etc., while having conflicting opinions, interests and preferences. Different perspectives can be drawn from this setting; we introduce what we think is interesting to do.

- Efficient tools for group preference elicitation. Most of the work on preference learning in MCDA focuses on representing the preferences of a single decision-maker (DM). In contrast, several real-world situations involve a group of decision-makers in the decision process. Therefore, a challenging question could be developing tools for group preference elicitation, allowing each group member to provide individual preference information to build a collective preference model accepted by each decision-maker. Different issues arise, among others: Which formal language (mathematical programming, Boolean formulation, etc.) can we rely on to build efficient algorithms? How to manage inconsistency and revision in this

setting?

- Multi-party dialogue: In the context of multi-agent systems, argumentation theory is a means to facilitate multi-agent interaction, as it naturally provides tools to design, implement and analyze sophisticated forms of interaction between rational agents. It provides a framework for structuring interaction between agents with potentially conflicting views while ensuring that the exchange respects certain principles (e.g., consistency of statements and discussions between participants). The idea here is to rely on tools of argumentation theory to analyze, structure, and formalize collective decision-making mechanisms to construct an informed joint decision [Bisquert et al., 2019]. Several works on multiparty dialogues in argumentation exist [Bonzon and Maudet, 2011; Dignum and Vreeswijk, 2003]. However, several questions remain open. For example, how to aggregate the opinions/preferences of participants? Several aggregation tools/models exist; it is a question of setting up an efficient and effective way of doing so. Another issue is how to consider the participants' arguments during the interaction. For example, participants do not necessarily present all their arguments simultaneously and may even hide particular arguments for various reasons. They may also form coalitions or have different roles during the discussion. So, what rules should be put in place to structure the dialogue? Questions related to aggregating different arguments from different participants during the dialogue are also an issue [Coste-Marquis et al., 2007].

- Explainability for Collective Decision: In this case, we want to do the same work we have done in defining argument schemes for decisions. These schemes took into account a decision-maker's preferences and features of the decision model. We will try to see to what extent we can extend our work to a context with several participants in the decision process. For instance, how can we ensure that the participants accept the final decision? For example, it is a question of extracting sufficient reasons that will support the joint decision, allowing the adoption of this decision by the participants. Working in this direction will be an opportunity to collaborate with colleagues in the Social Choice field, especially Anaëlle Wilczynski (MICS, CentraleSupélec).

5.5   Summary

This chapter has exposed our ambitions for the next years and the research questions we envisage answering to contribute to the Artificial intelligence and Decision theory fields. The different questions will offer us great opportunities to collaborate with various colleagues and future PhD students. We mentioned different possible new collaborations, but our actual collaborations will continue without any doubt and with

much pleasure.

We also have other projects that are not detailed in this manuscript. These projects reflect our desire to, on the one hand, enrich our scientific background and, on the other hand, to mobilize our knowledge acquired over the last years in new fields and challenges in collaboration with some colleagues. As examples, we mention the following two theses, where we will have the chance to participate in the supervision.

- Angélique Yameogo (October 2022). An XAI approach for the characterization, Conceptualization and Detection of Fake News. Co-supervision with Régis Fleurquin (IRISA, UMR CNRS 6074, Université de Bretagne Sud) and Nicolas Belloir (CREC St-Cyr, IRISA, UMR CNRS 6074, Université de Bretagne Sud). In collaboration also with Oscar Pastor (PROSS, Universidad Politécnica de Valencia, Spain).

- Dao Thauvin (November 2022). Explanatory dialogue for the interpretation of visual scenes [4]. Co-supervision with Stéphane Herbin (ONERA[5], the French Aerospace Lab) and Céline Hudelot (MICS, CentraleSupélec, Université Paris-Sacaly).

---

[4]In french: Dialogue explicatif pour l'interprétation de scènes visuelles.
[5]https://www.onera.fr/en/identity

Bibliography

Loïc Adam and Sébastien Destercke. Possibilistic Preference Elicitation by Minimax Regret. In *37th Conference on Uncertainty in Artificial Intelligence (UAI 2021)*, volume 161, pages 718–727, Online, United States, 2021. (Cited on page 88.)

Jose M Alonso and A Bugarın. Expliclas: Automatic generation of explanations in natural language for weka classifiers. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 660–665. IEEE, 2019. (Cited on page 86.)

Leila Amgoud and Henri Prade. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3-4):413–436, 2009. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/j.artint.2008.11.006. (Cited on page 77.)

Leila Amgoud, Nicolas Maudet, and Simon Parsons. Modelling dialogues using argumentation. In *Proceedings Fourth International Conference on MultiAgent Systems*, pages 31–38, 2000. (Cited on page 77.)

Manuel Amoussou. *Explication interactives dans l'aide à la décision multicriyère: gestion des inconssitences et des niveaux d'explication.* PhD thesis, CentraleSupélec, Université Paris Saclay, (in progress). (Cited on pages 48, 50, 65, 76 and 83.)

Manuel Amoussou, Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Explaining Robust Additive Decision Models: Generation of Mixed Preference-Swaps by Using MILP. In *From Multiple Criteria Decision Aid to Preference Learning (DA2PL 2020)*, Trento (virtual), Italy, 2020. (Cited on page 65.)

Manuel Amoussou, Khaled Belahcene, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Des explications par étapes pour le modèle additif. Journées d'Intelligence Artificielle Fondamentale, 2022. (Cited on page 62.)

Abdallah Arioua, Madalina Croitoru, and Patrice Buche. DALEK: a Tool for Dialectical Explanations in Inconsistent Knowledge Bases. In *COMMA: Computational Models of Argument*, volume 287, pages 461–462. IOS Press, 2016. (Cited on page 90.)

Abdallah Arioua, Patrice Buche, and Madalina Croitoru. Explanatory dialogues with argumentative faculties over inconsistent knowledge bases. *Expert Systems with Applications*, 80:244–262, 2017. (Cited on page 90.)

Katie Atkinson and Trevor Bench-Capon. Argumentation schemes in ai and law. *Argument & Computation*, 12:1–18, 03 2021. (Cited on page 49.)

Katie Atkinson, Trevor J. M. Bench-Capon, and Peter McBurney. A dialogue game protocol for multi-agent argument over proposals for action. *Auton. Agents Multi Agent Syst.*, 11(2):153–171, 2005. (Cited on page 77.)

Katie Atkinson, Trevor J. M. Bench-Capon, and Sanjay Modgil. Argumentation for decision support. In Stéphane Bressan, Josef Küng, and Roland R. Wagner, editors, *Database and Expert Systems Applications, 17th International Conference, DEXA 2006, Kraków, Poland, September 4-8, 2006, Proceedings*, volume 4080 of *Lecture Notes in Computer Science*, pages 822–831. Springer, 2006. (Cited on page 77.)

Ismaïl Baaj and Jean-Philippe Poli. Natural language generation of explanations of fuzzy inference decisions. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 563–568. IEEE, 2019. (Cited on pages 86 and 87.)

Ismaïl Baaj, Jean-Philippe Poli, and Wassila Ouerdane. Some insights Towards a Unified Semantic Representation of Explanation for eXplainable Artificial Intelligence (XAI). In *The 1st workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI)*, Tokyo, Japan, 2019. (Cited on page 86.)

Ismaïl Baaj, Jean-Philippe Poli, Wassila Ouerdane, and Nicolas Maudet. Representation of Explanations of Possibilistic Inference Decisions. In *ECSQARU 2021: European Conference on Symbolic and Quantitative Approaches with Uncertainty*, volume 12897 of *Lecture Notes in Computer Science*, pages 513–527, Prague, Czech Republic, 2021. Springer. (Cited on page 48.)

Ismaïl Baaj. *Explainability of Possibilistic and Fuzzy rule-based systems*. PhD thesis, Sorbonne Université, 2022. (Cited on pages 6, 48 and 86.)

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. (Cited on pages 5 and 48.)

Khaled Belahcene. *Towards accountable decision aiding: explanations for the aggregation of preferences*. PhD thesis, CentraleSupélec, Université Paris-Saclay, 2018. (Cited on pages 31, 48 and 50.)

Khaled Belahcene, Chirstophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183, 2017a. (Cited on pages 6, 51, 62, 64 and 65.)

Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. A model for accountable ordinal sorting. In *Proceedings of the 26th IJCAI*, pages 814–820, 2017b. (Cited on pages 6, 51 and 83.)

Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. An efficient SAT formulation for learning multiple criteria non-compensatory sorting rules from examples. *Computers & Operations Research*, 97: 58–71, 2018a. (Cited on pages 9, 28, 31, 32 and 35.)

Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Accountable approval sorting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*, 2018b. (Cited on pages 6, 28, 29, 32, 33, 51, 65, 68, 70, 72 and 92.)

Khaled Belahcene, Vincent Mousseau, Wassila Ouerdane, Marc Pirlot, and Olivier Sobrie. Ranking with Multiple reference Points-Efficient Elicitation and Learning Procedure. In *Proceeding of the 4th wokshop from multiple criteria Decision aid to Preference Learning (DA2PL)*, 2018c. (Cited on pages 9, 28 and 46.)

Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Comparing options with argument schemes powered by cancellation. In *Proceedings of IJCAI-19*, pages 1537–1543. International Joint Conferences on Artificial Intelligence Organization, 2019. (Cited on pages 6, 51 and 62.)

Khaled Belahcene, Vincent Mousseau, Wassila Ouerdane, Marc Pirlot, and Olivier Sobrie. Ranking with multiple points: Efficient elicitation and learning procedures. Submitted to Computers & OR, 2022. (Cited on page 46.)

Khaled Belahcène, Vincent Mousseau, Wassila Ouerdane, Marc Pirlot, and Olivier Sobrie. Ranking with multiple reference points: Efficient sat-based learning procedures. *Computers & Operations Research*, 150:106054, 2023. (Cited on page 28.)

Nawal Benabbou, Patrice Perny, and Paolo Viappiani. Incremental elicitation of choquet capacities for multicriteria choice, ranking and sorting problems. *Artif. Intell.*, 246: 152–180, 2017. doi: 10.1016/j.artint.2017.02.001. URL https://doi.org/10.1016/j.artint.2017.02.001. (Cited on pages 18 and 88.)

Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artif. Intell.*, 171(10-15):619–641, 2007. (Cited on page 77.)

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, page 1, 2017. (Cited on pages 5 and 48.)

Pierre Bisquert, Madalina Croitoru, Christos Kaklamanis, and Nikos Karanikolas. A Decision-Making approach where Argumentation added value tackles Social Choice deficiencies. *Progress in Artificial Intelligence*, 8(2):229–239, 2019. (Cited on page 93.)

Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34, 1948. (Cited on pages 11 and 28.)

Duncan Black. *The theory of committees and elections*. University Press, Cambridge, 1958. (Cited on pages 8, 11 and 28.)

Elizabeth Black, Nicolas Maudet, and Simon Parsons. Argumentation-based Dialogue. In Dov Gabbay, Massimiliano Giacomin, Guillermo R. Simari, and Matthias Thimm, editors, *Handbook of Formal Argumentation, Volume 2*. College Publications, 2021. URL https://hal.archives-ouvertes.fr/hal-03429859. (Cited on pages 10, 76 and 77.)

Bart Bogaerts, Emilio Gamba, and Tias Guns. A framework for step-wise explaining how to solve constraint satisfaction problems. *Artif. Intell.*, 300:103–550, 2021. (Cited on page 60.)

Arthur Boixel, Ulle Endriss, and Ronald de Haan. A calculus for computing structured justifications for election outcomes. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI-2022)*, February 2022. (Cited on page 60.)

Elise Bonzon and Nicolas Maudet. On the outcomes of multiparty persuasion. In *Argumentation in Multi-Agent Systems. Eighth International Workshop, ArgMAS 2011. Revised, Selected and Invited Papers.*, volume 7543 of *Lecture Notes in Computer Science*, pages 86–101, Taipei, Taiwan, May 2011. Springer. (Cited on page 93.)

Craig Boutilier and Jeffrey S. Rosenschein. *Incomplete Information and Communication in Voting*, page 223–258. Cambridge University Press, 2016. (Cited on page 72.)

Denis Bouyssou. Outranking methods. In C. A. Floudas and P. M. Pardalos, editors, *Encyclopedia of Optimization*, pages 2887–2893. Springer, 2009. ISBN 978-0-387-74758-3. (Cited on page 13.)

Denis Bouyssou and Thierry Marchant. An axiomatic approach to noncompensatory sorting methods in MCDM, I: the case of two categories. *European Journal of Operational Research*, 178(1):217–245, 2007a. (Cited on pages 6, 20, 21, 24 and 92.)

Denis Bouyssou and Thierry Marchant. An axiomatic approach to noncompensatory sorting methods in MCDM, II: more than two categories. *European Journal of Operational Research*, 178(1):246–276, 2007b. (Cited on pages 6, 20, 21 and 24.)

Denis Bouyssou and Thierry Marchant. Multiattribute preference models with reference points. *European Journal of Operational Research*, 229(2):470 – 481, 2013. (Cited on page 45.)

Denis Bouyssou, Thierry Marchant, Patrice Perny, Marc Pirlot, Alexis Tsoukiàs, and Philippe Vincke. *Evaluation and decision models: a critical perspective*, volume 32 of *International Series in Operations Research and Management Science*. Kluwer Academic Publishers, 2000. (Cited on pages 2, 16 and 17.)

Denis Bouyssou, Thierry Marchant, Marc Pirlot, Alexis Tsoukiàs, and Philippe Vincke. *Evaluation and decision models with multiple criteria: Stepping stones for the analyst.* Springer Verlag, Boston, 2006. (Cited on pages 1, 2, 13 and 18.)

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. *Handbook of Computational Social Choice.* Cambridge University Press, USA, 1st edition, 2016. (Cited on page 92.)

Olivier Cailloux and Ulle Endriss. Arguing about voting rules. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 287–295. ACM, 2016. (Cited on page 60.)

Giuseppe Carenini and Johanna D. Moore. Generating and evaluating evaluative arguments. *Artificial Intelligence Journal*, 170:925–952, 2006. (Cited on pages 5 and 48.)

Lauri Carlson. *Dialogue Games: An Approach to Discourse Analysis.* Reidel, 1983. (Cited on page 5.)

Michael Cashmore, Anna Collins, Benjamin Krarup, Senka Krivic, Daniele Magazzeni, and David Smith. Towards Explainable AI Planning as a service. In *International Conference on Automated Planning and Scheduling second workshop on Explainable Planning*, 2019. doi: 10.48550/arxiv.1908.05059. (Cited on page 48.)

J. Arturo Castillo-Salazar, Dario Landa-Silva, and Rong Qu. Workforce scheduling and routing problems: literature survey and computational study. *Annals of Operations Research*, 239(1):39 – 67, 2016. (Cited on page 84.)

Alison Cawsey. Planning interactive explanations. *International Journal of Man-Machine Studies*, 38(2):169–199, 1993. ISSN 0020-7373. (Cited on page 85.)

Balakrishnan Chandrasekaran, Michael Tanner, and John Josephson. Explaining control strategies in problem solving. *IEEE Expert*, 4:9–15, 1989. doi: 10.1109/64.21896. (Cited on page 48.)

Shruthi Chari, Daniel M. Gruen, Oshani Seneviratne, and Deborah L. McGuinness. Directions for explainable knowledge-enabled systems. In *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, pages 245–261. Ios Press, 2020. (Cited on page 86.)

Yasmine Charif-Djebbar and Nicolas Sabouret. An agent interaction protocol for ambient intelligence. In *2006 2nd IET International Conference on Intelligent Environments - IE 06*, volume 1, pages 275–284, 2006. (Cited on page 77.)

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47:547–553, 2009. (Cited on page 44.)

Sylvie Coste-Marquis and Pierre Marquis. From Explanations to Intelligible Explanations. In *1st International Workshop on Explainable Logic-Based Knowledge Representation (XLoKR'20)*, Rhodes, Greece, 2020. Workshop at KR'20. (Cited on page 7.)

Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex, and Pierre Marquis. On the merging of dung's argumentation systems. *Artificial Intelligence*, 171(10):730–753, 2007. (Cited on page 93.)

Dimitiris K. Despotis and Constantin Zopounidis. *Building Additive Utilities in the Presence of Non-Monotonic Preferences*, pages 101–114. Springer, 1995. (Cited on page 41.)

Sébastien Destercke. A generic framework to include belief functions in preference handling and multi-criteria decision. *International Journal of Approximate Reasoning*, 98:62–77, 2018. (Cited on page 90.)

Frank Dignum and Gerard Vreeswijk. Towards a testbed for multi-party dialogues. In *Advances in Agent Communication, International Workshop on Agent Communication Languages, ACL*, pages 212–230, 2003. (Cited on page 93.)

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. (Cited on pages 5 and 48.)

Michael Doumpos and Constantin Zopounidis. Preference disaggregation and statistical learning for multicriteria decision support: A review. *European Journal of Operational Research*, 209(3):203 – 214, 2011. (Cited on page 18.)

Karim El Mernissi. *Une étude de la génération d'explication dans un système à base de règles*. PhD thesis, Univeristé Pierre et Marie Curie, 2017. (Cited on pages 1, 6 and 48.)

Fredrik Engström and Claes Strannegård Abdul Rahim Nizamani. Generating Comprehensible Explanations in Description Logic. In *Informal Proceedings of the 27th International Workshop on Description Logics*, Vienna, 2014. (Cited on page 60.)

George Ferguson, James Allen, and Brad Miller. Trains-95: Towards a mixed-initiative planning assistant. In *Proceedings of the 3rd. International Conference on AI Planning Systems*, 1996. (Cited on page 5.)

Valentina Ferretti, Jinyan Liu, Vincent Mousseau, and Wassila Ouerdane. Reference-based ranking procedure for environmental decision making: Insights from an ex-post analysis. *Environmental Modelling & Software*, 99:11 – 24, 2018. (Cited on page 45.)

José Rui Figueira, Vincent Mousseau, and Bernard Roy. Electre methods. In *Multiple criteria decision analysis: State of the art surveys*, pages 133–153. Springer, 2005. (Cited on page 36.)

Peter C. Fishburn. Condorcet social choice functions. *SIAM Journal on Applied Mathematics*, 33(3):469–489, 1977. (Cited on page 54.)

James Forrest, Somayajulu Sripada, Wei Pang, and George Coghill. Towards making nlg a voice for interpretable machine learning. In *Proceedings of The 11th International Natural Language Generation Conference*, pages 177–182, 2018. (Cited on page 86.)

John Fox and Simon Parsons. Arguing about beliefs and actions. In *Applications of uncertainty formalisms*. Springer-Verlag, 1998. (Cited on page 77.)

Johanne Furnkranz and Eyke Hullermeier. *Preference Learning*. Springer, 2011. ISBN 978-3-642-14124-9. (Cited on pages 8, 14, 19, 27 and 29.)

Albert Gatt and Ehud Reiter. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, 2009. (Cited on page 86.)

Hugo Gilbert, Nawal Benabbou, Patrice Perny, Olivier Spanjaard, and Paolo Viappiani. Incremental decision making under risk with the weighted expected utility model. In *Proceedings of the 26 International Joint Conference on Artificial Intelligence*, pages 4588–4594, 2017. (Cited on page 88.)

Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018. (Cited on page 5.)

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017. (Cited on page 47.)

Michel Grabisch and Christophe Labreuche. A decade of application of the choquet and sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175(1):247–286, 2010. (Cited on page 17.)

Salvatore Greco, Roman Słowiński, José Rui Figueira, and Vincent Mousseau. *Robust Ordinal Regression*, pages 241–283. Springer US, 2010. (Cited on pages 54, 65 and 72.)

Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4), 1999. (Cited on pages 5 and 48.)

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2019. (Cited on pages 5 and 48.)

Adel Guitouni and Jean-Marc Martel. Tentative guidelines to help choosing an appropriate MCDA method. *European Journal of Operational Research*, 109(2):501–521, 1998. (Cited on page 17.)

David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017. (Cited on pages 1, 5 and 47.)

Sharmi Dev Gupta, Begum Genc, and Barry O'Sullivan. Finding counterfactual explanations through constraint relaxations, 2022. URL https://arxiv.org/abs/2204.03429. (Cited on page 48.)

John S. Hammond, Ralph L. Keeney, and Howard Raiffa. Even swaps: A rational method for making trade-offs. *Harvard business review*, 76:137–8, 143, 03 1998. (Cited on pages 61 and 64.)

Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pages 241–250, 2000. (Cited on pages 5 and 48.)

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: challenges and prospects. *CoRR*, abs/1812.04608, 2018. (Cited on page 87.)

James Hollan, Edwin Hutchins, and David Kirsh. Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.*, 7(2):174—-196, 2000. (Cited on page 86.)

Matthew Horridge, Samantha Bail, Bijan Parsia, and Uli Sattler. Toward cognitive support for OWL justifications. *Know.-Based Syst.*, 53:66–79, nov 2013. ISSN 0950-7051. (Cited on page 60.)

Eric Horvitz. Uncertainty, action, and interaction: In pursuit of mixed-initiative computing. *Intelligent Systems*, pages 17–20, 2000. (Cited on page 88.)

Eyke Hüllermeier. Preference learning: Machine learning meets MCDA. In *DA2PL 2014 Workshop From Multiple Criteria Decision Aid to Preference Learning*, pages 1–2, 2014. Paris, France. (Cited on page 8.)

Eric Jacquet-Lagrèze and Yannis Siskos. Preference disaggregation: 20 years of MCDA experience. *European Journal of Operational Research*, 130(2):233–245, 2001. (Cited on page 8.)

Ulrich Junker. Quickxplain: Preferred explanations and relaxations for over-constrained problems. In *Proceedings of the 19th National Conference on Artifical Intelligence*, pages 167–172, 2004. (Cited on pages 56, 72 and 90.)

Souhila Kaci. *Working with Preferences: Less Is More.* Cognitive Technologies. Springer, 2011. (Cited on pages 8 and 27.)

Milosz Kadzinski, Salvatore Greco, and Roman Slowinski. Selection of a representative value function in robust multiple criteria ranking and choice. *European Journal of Operational Research*, 217(3):541 – 553, 2012. (Cited on page 19.)

Milosz Kadziński and Krzystof Ciomek. Active learning strategies for interactive elicitation of assignment examples for threshold-based multiple criteria sorting. *Eur. J. Oper. Res.*, 293(2):658–680, 2021. (Cited on page 18.)

Anotni Kakas and Pavlos Moraitis. Argumentation based decision making for autonomous agents. In *Proc. AAMAS*, 2003. (Cited on page 77.)

Ralph L. Keeney and Howard Raiffa. *Decisions with multiple objectives: Preferences and value tradeoffs.* J. Wiley, New York, 1976. (Cited on page 19.)

John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959. (Cited on page 45.)

Arwa Khannoussi, Alexandru Liviu Olteanu, Christophe Labreuche, Pritesh Narayan, Catherine Dezan, Jean-Philippe Diguet, Jacques Petit-Frère, and Patrick Meyer. Integrating operators' preferences into decisions of unmanned aerial vehicles: Multi-layer decision engine and incremental preference elicitation. In *Algorithmic Decision Theory, Proceedings*, volume 11834, pages 49–64, 2019. (Cited on page 45.)

David H. Krantz, R.Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of measurement*, volume 1: Additive and Polynomial Representations. Academic Press, 1971. (Cited on page 62.)

Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2017. (Cited on page 67.)

Christophe Labreuche, Nicolas Maudet, and Wassila Ouerdane. Minimal and complete explanations for critical multi-attribute decisions. In *ADT*, pages 121–134, 2011. (Cited on pages 6, 51 and 54.)

Christophe Labreuche, Nicolas Maudet, and Wassila Ouerdane. Justifying dominating options when preferential information is incomplete. In *ECAI 2012.*, pages 486–491, 2012. (Cited on pages 6, 51, 55, 56 and 62.)

Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane, and Simon Parsons. A dialogue game for recommendation with adaptive preference models. In *Proceedings AAMAS*, pages 959–967, 2015. (Cited on pages 10, 77, 78, 79, 82 and 83.)

Jean-François Laslier and M. Remzi Sanver. *Handbook on Approval Voting.* Studies in Choice and Welfare. Springer, Boston, 2010. (Cited on page 66.)

Mohammed El Amine Lazouni, Mohammed Amine Chikh, and Mahmoudi Said. A new computer aided diagnosis system for pre-anesthesia consultation. *Journal of Medical Imaging and Health Informatics*, 3(4):471–479, 2013. (Cited on page 44.)

Mathieu Lerouge. *Conception de méthodes d'explication des résultats obtenus par des systèmes d'optimisation : application à des problèmes de planification.* PhD thesis, CentraleSupélec, Université Paris Saclay, (in progress). (Cited on pages 1, 6, 48 and 84.)

Agnes Leroy, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a multiple criteria sorting method. In *International Conference on Algorithmic Decision Theory*, pages 219–233. Springer, 2011. (Cited on pages 8, 9, 24, 28, 30 and 41.)

Qingzi Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing design practices for Explainable AI User Experiences. In *Proceedings of the 2020 Conference on Human Factors in Computing Systems*, page 1 – 15. Association for Computing Machinery, 2020. (Cited on page 48.)

Christopher J. Lindsell, William W. Stead, and Kevin B. Johnson. Action-Informed Artificial Intelligence—Matching the Algorithm to the Problem. *JAMA*, 323(21): 2141–2142, 2020. (Cited on page 85.)

Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27: 247–266, 1990. (Cited on pages 48 and 84.)

Jinyan Liu. *Preference elicitation for multi-criteria ranking with multiple reference points.* PhD thesis, CentraleSupélec, Université Paris-Saclay, 2016. (Cited on page 45.)

Jinyan Liu, Wassila Ouerdane, and Vincent Mousseau. A metaheuristic approach for preference learning in multicriteria ranking based on reference points. In *Proceedings of the 2nd workshop "From multiple criteria Decision Aid to Preference Learning" (DA2PL)*, pages 76–86, Chatenay-Malabry, France, 2014. (Cited on pages 9, 28 and 45.)

Manel Maamar. *Modélisation et optimisation bi-objectif et multi-période avec anticipation d'une place de marché de prospects Internet : adéquation offre/demande.* Theses, Université Paris Saclay, 2015. (Cited on page 1.)

Massinissa Mammeri. *Decision aiding methodology for developing the Contractual Strategy of complex oil and gas development projects.* Theses, Université Paris-Saclay, 2017. (Cited on page 1.)

Peter McBurney and Simon Parsons. Dialogue game protocols. *Agent Communication Languages*, pages 269–283, 2003. (Cited on pages 5 and 10.)

John McClure. Goal-based explanations of actions and outcomes. *European Review of Social Psychology*, 12(1):201–235, 2002. (Cited on page 87.)

Lorraine McGinty and Barry Smyth. Adaptive selection: An analysis of critiquing and preference-based feedback in conversational recommender systems. *International Journal of Electronic Commerce*, 11(2):35–57, 2006. (Cited on page 9.)

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38, 2019. (Cited on pages 5, 7, 48, 87 and 88.)

Pegdwende Minoungou. *Apprentissage de modèles à règle majoritaire à partir de données partiellement monotones.* PhD thesis, CentraleSupélec, Université Paris Saclay, 2022. (Cited on pages 39 and 44.)

Pegdwendé Minoungou, Vincent Mousseau, Wassila Ouerdane, and Paolo Scotton. Learning an MR-sort model from data with latent criteria preference direction. In *The 5th workshop from multiple criteria Decision Aid to Preference Learning (DA2PL)*, 2020. (Cited on pages 9, 28, 39 and 40.)

Pegdwendé Minoungou, Vincent Mousseau, Wassila Ouerdane, and Paolo Scotton. A MIP-based approach to learn MR-Sort models with single-peaked preferences. *Annals of Operations Research*, 2022. (Cited on pages 9, 28 and 43.)

Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*, 2018. (Cited on pages 5 and 48.)

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 2021. (Cited on pages 48 and 87.)

Vincent Mousseau, L.C. Dias, J. Figueira, C. Gomes, and J.N. Clímaco. Resolving inconsistencies among constraints on the parameters of an MCDA model. *European Journal of Operational Research*, 147(1):72–93, 2003. (Cited on page 90.)

Jann Müller and Anthony Hunter. An argumentation-based approach for decision making. In *Proc. ICTAI*, 2012. (Cited on page 77.)

Ingrid Nunes, Simon Miles, Michael Luck, Simone Barbosa, and Carlos Lucena. Pattern-based explanation for automated decisions. In *Proceedings of the 21st ECAI*, pages 669–674. IOS Press, 2014. (Cited on pages 5 and 48.)

Alexandru Liviu Olteanu, Khaled Belahcene, Vincent Mousseau, Wassila Ouerdane, Aantoine Rolland, and June Zheng. Preference elicitation for a ranking method based on multiple reference profiles. *4OR: A Quarterly Journal of Operations Research*, 2021. to appear. (Cited on pages 9 and 28.)

Wassila Ouerdane. *Multiple criteria decision aiding : a dialectical perspective*. PhD thesis, Université Paris Dauphine - Paris IX, 2009. (Cited on pages 1, 10, 48, 50, 76 and 77.)

Wassila Ouerdane, Nicolas Maudet, and Alexis Tsoukiàs. Argument schemes and critical questions for decision aiding process. In Ph. Besnard, S. Doutre, and A. Hunter, editors, *Proceedings of the 2nd International Conference on Computational Models of Argument(COMMA'08)*, pages 285–296, 2008. (Cited on pages 10, 77 and 82.)

Wassila Ouerdane, Nicolas Maudet, and Alexis Tsoukiàs. Dealing with the dynamics of proof-standard in argumentation-based decision aiding. In *Proc. ECAI*, pages 999–1000, 2010. (Cited on pages 10, 77, 79 and 82.)

Wassila Ouerdane, Yannis Dimopoulos, Konstantinos Liapis, and Pavlos Moraitis. Towards automating Decision Aiding through Argumentation. *Journal of Multi-Criteria Decision Analysis*, 18(5-6):289–309, 2011. (Cited on pages 10, 77 and 82.)

Meltem Ozturk, Alexis Tsoukias, and Philippe Vincke. Preference Modelling. In J. Figueira, S. Greco, , and M. Ehrgott, editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 27–72. Springer Verlag, Boston, Dordrecht, London, 2005. (Cited on page 27.)

Bart Peintner, Paolo Viappiani, and Neil Yorke-Smith. Preferences in interactive systems: Technical challenges and case studies. *AI Magazine*, 29(4):13, Dec. 2008. (Cited on pages 8 and 45.)

Patrice Perny. *Modélisation des préférences, agrégation multicritère et systèmes d'aide à la décision*. PhD thesis, Mémoire présenté en vue de l'obtention de l'habilitation à diriger des recherches, Université Pierre et Marie Curie, 2000. (Cited on page 17.)

Patrice Perny, Paolo Viappiani, and Abdellah Boukhatem. Incremental preference elicitation for decision making under risk with the rank-dependent utility model. In Alexander T. Ihler and Dominik Janzing, editors, *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016*. AUAI Press, 2016. (Cited on page 88.)

Philip Pettit. *Republicanism: A Theory of Freedom and Government*. Oxford University Press, 1997. (Cited on page 67.)

Philip Pettit. Democracy, electoral and contestatory. In Ian Shapiro and Stephen Macedo, editors, *Designing Democratic Institutions*, pages 105–144. New York, USA: New York University Press, 2000. (Cited on page 67.)

Régis Pierrard, Jean-Philippe Poli, and Céline Hudelot. A new approach for explainable multiple organ annotation with few data. In *Proceedings of the Workshop on Explainable Artificial Intelligence (XAI) 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, XAI@IJCAI 2019*, pages 107–113, 2019. (Cited on page 86.)

Gabriella Pigozzi, Alexis Tsoukiàs, and Paolo Viappiani. Preferences in Artificial Intelligence. *Annals of Mathematics and Artificial Intelligence*, 77(3-4):361–401, 2016. (Cited on page 8.)

Vladislav V. Podinovskii. Criteria importance theory. *Mathematical Social Sciences*, 27 (3):237–252, 1994. (Cited on page 18.)

Jean-Philippe Poli, Wassila Ouerdane, and Régis Pierrard. Generation of textual explanations in xai: the case of semantic annotation. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2021. (Cited on pages 48 and 87.)

Ariel D. Procaccia. Axioms should explain solutions. *The Future of Economic Design*, 2019. (Cited on page 60.)

Stephen J. Read and Amy Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65:429–447, 1993. (Cited on page 87.)

Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Natural Language Processing. Cambridge University Press, 2000. (Cited on page 86.)

Aantoine Rolland. *Procédures d'agrégation ordinale de préférences avec points de référence pour l'aide à la décision*. PhD thesis, Université Paris 6, France, 2008. (Cited on page 45.)

Antoine Rolland. Reference-based preferences aggregation procedures in multi-criteria decision making. *European Journal of Operational Research*, 225(3):479 – 486, 2013. (Cited on pages 8, 17, 28 and 45.)

Bernard Roy. The outranking approach and the foundations of Electre methods. *Theory and Decision*, 31(1):49–73, 1991. (Cited on pages 13 and 20.)

Bernard Roy. *Multicriteria Methodology for Decision Aiding*. Kluwer Academic, Dordrecht, 1996. (Cited on page 1.)

Bernard Roy and Vincent Mousseau. A theoretical framework for analysing the notion of relative importance of criteria. *Journal of Multi-Criteria Decision Analysis*, 5: 145–159, 1996. (Cited on page 18.)

Bernard Roy and R. Słowiński. Questions guiding the choice of a multicriteria decision aiding method. *EURO Journal on Decision Processes*, 1(1):69–97, 2013. (Cited on page 17.)

Guillermo Ricardo Simari and Iyad Rahwan, editors. *Argumentation in Artificial Intelligence*. Springer, 2009. ISBN 978-0-387-98196-3. (Cited on page 77.)

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. (Cited on page 48.)

Yannis Siskos, Evangelos Grigoroudi, and Nikolaos F. Matsatsinis. Uta methods. In *Multiple criteria decision analysis: State of the art surveys*, pages 297–334. Springer, 2005. (Cited on page 19.)

Olivier Sobrie. *Learning preferences with multiple-criteria models*. PhD thesis, Université de Mons (Faculté Polytechnique) and Université Paris-Saclay (CentraleSupélec), June 2016. (Cited on pages 39 and 40.)

Olivier Sobrie, Vincent Mousseau, and M. Pirlot. Learning a Majority Rule Model from Large Sets of Assignment Examples. In *Algorithmic Decision Theory*, volume 8176, pages 336–350. Berlin, Heidelberg, 2013. (Cited on page 44.)

Olivier Sobrie, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a non compensatory sorting model. In *Algorithmic Decision Theory*, volume 9346, pages 153–170. Springer, 2015. (Cited on page 30.)

Olivier Sobrie, Vincent Mousseau, and M. Pirlot. Learning monotone preferences using a majority rule sorting model. *International Transactions in Operational Research*, 26(5):1786–1809, 2019. (Cited on pages 30, 39 and 40.)

William Swartout. Xplain: A system for creating and explaining expert consulting programs. *Artif. Intell.*, 21:285–325, 1983. (Cited on pages 5 and 48.)

William Swartout and Stephen Smoliar. On making expert systems more like experts. *Expert Systems*, 4(3):196–208, 1987. doi: 10.1111/j.1468-0394.1987.tb00143.x. (Cited on page 48.)

Paul Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–467, 1989. doi: 10.1017/S0140525X00057046. (Cited on page 87.)

Nina Tintarev. Explanations of recommendations. In *Proc. ACM conference on Recommender systems*, pages 203–206, 2007. (Cited on pages 5 and 48.)

Ali Tlili. *Modèles de tri contraint multicritères pour la sélection de portefeuilles*. PhD thesis, CentraleSupélec, Université Paris Saclay, 2022. (Cited on page 1.)

Ali Tlili, Khaled Belahcène, Oumaima Khaled, Vincent Mousseau, and Wassila Ouerdane. Learning non-compensatory sorting models using efficient sat/maxsat formulations. *European Journal of Operational Research*, 298(3):979–1006, 2022. (Cited on pages 9, 28, 31, 32, 34, 35 and 36.)

Alexis Tsoukiàs. On the concept of decision aiding process. *Annals of Operations Research*, pages 3 – 27, 2007. (Cited on page 8.)

Alexis Tsoukiàs. From decision theory to decision aiding methodology. *European Journal of Operational Research*, 187:138–161, 2008. (Cited on pages 1, 2, 13 and 77.)

Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, 36: e5, 2021. (Cited on page 77.)

Paolo Viappiani and Craig Boutilier. Regret-based optimal recommendation sets in conversational recommender systems. In *Proceedings of the Third ACM Conference on Recommender Systems*, page 101–108, 2009. (Cited on page 83.)

Paolo Viappiani, Boi Faltings, and Pearl Pu. Preference-based search using example-critiquing with suggestions. *J. Artif. Int. Res.*, 27(1):465–503, dec 2006. ISSN 1076-9757. (Cited on page 9.)

Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021. ISSN 1566-2535. (Cited on page 5.)

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2):76–99, 2017. (Cited on page 47.)

Peter Wakker. *Additive Representations of Preferences: A New Foundation of Decision Analysis*. Theory and Decision Library C. Springer Netherlands, 1989. (Cited on page 62.)

Douglas Walton. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996. (Cited on pages 7, 11, 49, 60, 73 and 81.)

Douglas N. Walton and Eric C.W. Krabbe. *Commitment in Dilaogue : Basic conceptions of Interpersonal Reasoning*. State University of New York Press, 1995. (Cited on pages 5, 76 and 77.)

Donald Waterman. *A guide to expert systems*. Addison-Wesley Pub. Co., Reading, MA, 1986. (Cited on page 17.)

Michael Wick and William Thompson. Reconstructive Expert System explanation. *Artificial Intelligence*, 54:33–70, 1992. (Cited on page 48.)

Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative xai: A survey, 2021. URL https://arxiv.org/abs/2105.11266. (Cited on page 77.)

Appendices

Curriculum Vitae

Wassila OUERDANE

Assistant Professor in Computer Science

Artificial Intelligence & Decision Aid

November 2022

ADDRESS:	CentraleSupélec-Bâtiment Bouygues
	Laboratoire de Mathématiques et Informatique pour la Complexité et
	les Systèmes (MICS)
	3, rue Joliot Curie 91190, Gif-Sur-Yvettes
PHONE:	+33 1 75 31 66 78
EMAIL:	wassila.ouerdane@centralesupelec.fr
WEB SITE:	https://wassilaouerdane.github.io

EDUCATION

1 DECEMBER 2009	**PhD** in COMPUTER SCIENCE, Paris Dauphine University
	Title: "Multiple Criteria Decision Aiding : a Dialectical Perspective."
	Supervisors: Alexis Tsoukiàs (DR CNRS, LAMSADE, Paris Dauphine University) and Nicolas Maudet (Assistant Professor, LAMSADE, Paris Dauphine University) .
	Jury:
	Referees: Simon Parsons (PR, Brooklyn College NY), Patrice Perny (PR, Université Pierre et Marie Curie)
	Members: Leila Amgoud (CR, CNRS, Université Paul Sabatier), Sylvie Coste-Marquis (MCF, Université d'artois), Thierry Marchant (PR, Ghent University Belgium), Christophe Labreuche (invité,Thales)
SEPTEMBER 2005	**Master degree** in COMPUTER SCIENCE. Paris Dauphine University Title: "How to choose a process modeling tool in a process of capitalizing on knowledge? "
SEPTEMBER 2003	**Engineering degree** in COMPUTER SCIENCE. Mouloud Mammeri University (Algeria). Title: Implementation of the AODV Routing protocol for Ad hoc mobile networks under Network Simulator.

Academic Positions

March. 2019 -	**Assistant Professor** at CentraleSupélec, Computer Science. Mathematics and Informatics Lab (MICS).
Sept. 2010- Feb. 2019	**Assistant Professor** at CentraleSupélec, Computer Science. Industrial Engineering Lab (LGI).
Sept. 2009- Sept. 2010	**Teaching and Research Assistant** in Computer Science. Paris Dauphine University, France.
Sept. 2008- Sept. 2009	**Teaching and Research Assistant** in Computer Science. Paris Dauphine University, France.
Sept. 2005- Sept. 2008	**PhD Candidate** at LAMSADE. Paris Dauphine University, France.
Sept. 2005- Sept. 2008	**Teaching Assistant** at Paris Dauphine University, France.

Collective Responsibilities

Nationals	Co-leader of the Working Group "Explainability and Trust" of the French AI Research Group (GDR IA[1]), starting Fall 2022, with Sébastien Destercke (DR, Heudiasyc, UTC)
Locals	Co-leader of the third year (3A) of CentraleSupélec, AI training (arround 70 students), september2022– with Céline Hudelot (MICS, CentraleSupélec)
	Co-Responsible of the Project Activity in AI, first and second (1A/2A) years of CentraleSupélec (L3-M1)–160 étudiants, since 2019 with Jean-Phlippe Poli (CEA-List)
	Member of the CentraleSupélec Restricted Scientific Board since 2019
	Elected member of the Scientific Board of CentraleSupélec, (Representative of lecturers and similar staff), since 2019
	Elected member of the LGI laboratory council, SEPT. 2010–FEV. 2019.

RESEARCH TOPICS

Our research addresses questions related to knowledge representation and reasoning in the context of eXplainable AI (XAI). Our main motivations are designing and modeling adaptive decision support systems to construct and support justified automatic recommendations. Our research lies at the intersection of the fields of Multi-Criteria Decision Aiding (MCDA) and Artificial Intelligence (knowledge representation and reasoning).

Multi-Criteria Decision Aiding (MCDA) aims to develop decision models explicitly based on the construction of a set of criteria reflecting the relevant aspects of the decision-making problem. These n criteria (often conflicting) ($\mathcal{N} = \{1, 2, \ldots, n\}$ with $n \geq 2$) evaluate a set of alternatives $A = \{a, b, c, \ldots\}$ from different points of view. Several multi-criteria decision models exist. These models correspond to a parametric family of functions aggregating the evaluation according to each criterion into a solution of the decision problem. The MCDA literature considers different decision problems. We distinguish the *choice*, the *sorting*, the *pairwise comparison*, and the *ranking*. Unlike formulations of choice, ranking and pairwise comparison problems, which are comparative, sorting formulates the decision problem in terms of assigning alternatives to predefined ordered categories $C^1, C^2, \ldots C^p$, where C^1 (C^p, resp.) is the worst (best, resp.) category. The assignment of an alternative to the appropriate category is based on its intrinsic value and not on its comparison with other alternatives.

In addition, multi-criteria decision aiding results from an interaction between at least two agents, an analyst and a decision-maker, where the analyst's goal is to guide the decision-maker in the construction and understanding of the recommendations of a particular decision problem. Decision theory and Multiple Criteria Decision Analysis (MCDA) have established the theoretical foundation upon which many decision support systems have risen. The different approaches (and the formal tools coming along with them) have focused for a long time on how a "solution" should be established. But it is clear that the process involves many other aspects that are handled more or less formally by the analyst. For instance,

- the problem of accountability of decisions is almost as important as the decision itself. The decision-maker should then be convinced by a proper explanation that the proposed solution is indeed the best.

- it should be possible, for the decision-maker, to refine, or even contradict, a given recommendation. Indeed, the decision-support process is often constructive, in the sense that the DM refines its formulation of the problem when confronted to potential solutions.

In addition, nowadays, decision support situations are omnipresent: they can arise when the analyst's role is assumed by a non-expert or even, in some cases, by an artificial agent. This means that several aspects - such as learning preferences, structuring the interaction, providing an explanation, handling user feedback,... - generally delegated to the human analyst should be ideally managed by the artificial agent. Thus, on the one hand, we need a formal theory on preferences and, on the other hand, a formal language making it possible to represent the dialogue and explain and communicate its results to convince the user that what is happening is both theoretically sound and operationally reasonable. In this context, the main (complementary) axes of my research work are:

Axis1: Modeling and generating explanations for recommendations for complex decision problems.

The question of the explanation (explainability/interpretability) of a decision, recommendation, algorithm outputs, etc., often associated in the literature with the acronym XAI (eXplainable AI), has become in recent years a crucial element in any "trusted algorithmic design". Indeed, for high-stakes AI applications, performance is not the only criterion to be taken into account. Such applications may require a relative understanding of the logic executed by the system. In this case, the end-user wants an answer to the question "Why?".

eXplainable Artificial Intelligence (XAI) aims to provide methods that help empower AIs to answer this question. Even though interest in this question has exploded with machine learning tools and techniques, it dates back to expert systems, and since then, many works have emerged. Various questions are explored, such as: generating and providing explanations, identifying desirable characteristics of an explanation from the point of view of its recipient, evaluating the explanation produced by the system, etc.

In general, my work focuses on *the implementation of tools and algorithms for generating explanations for recommendations stemming from multicriteria models* which put user preferences and judgments at the heart of the reasoning. Generating explanations in the MCDA context is not a simple task; as different criteria are at stake, the user cannot fully assess their importance or understand how they interact. Moreover, once the user is faced with the result and the explanation, he may realize that it is not exactly what he expected. Therefore, it can make changes or provide new information that will have effects, for example, on the other phases of the decision aiding process (e.g., preferences learning step, see Axis 2). Thus, beyond making the result acceptable, presenting an explanation can impact the representation of the user's reasoning mode, which is at the base of the construction of the recommendation. Furthermore, the challenge with this question is that the concept of explanation varies depending on the decision context/problem and the decision model. In this context, my research work focuses on two decision models: one very widely used model, whether in decision theory or in machine learning, namely the additive model, and the other which is Non-Compensatory Sorting model. With the first model, the work aims to produce explanations for the pairwise comparison. In contrast, in the second, we seek to explain the assignment of an alternative to a given category. To answer these questions, different approaches and techniques are considered: argumentation schemes and mathematical programming. In particular, the question of constructing explanations comes down to formalizing argument (explanations) schemes that link premises (information provided or approved by the user, or deduced during the process of preference learning, and some additional hypotheses on the process of reasoning (from the assumptions of the model)) to a conclusion (e.g. the recommendation) Finally, I am also interested in other models/systems, for example, rule-based systems (classical, fuzzy) and optimization models.

- **Concerned thesis**: Manuel Amoussou (in progress), Mathieu Lerouge (in progress), Ismail Baaj (2022), Khaled Belahcène (2018), Karim El Mernissi (2017).

Axis2: modelling of the interaction and preferences for the construction of adaptive decision support systems.

At present, when decision aiding support or recommendation systems (online, for example) are in full expansion, an important aspect is that of succeeding in capturing and integrating the preferences, habits, and reactions of users to try to produce the most compelling and relevant recommendations from a user perspective. To meet this objective, I investigated two lines of research.

- **Setting up efficient preference learning mechanisms**: learning and eliciting preferences is an essential step in a decision support process. This step aims to incorporate user judgments as faithfully as possible into the decision model. It is crucial to develop relevant and reliable recommendations, and any flawed process would lead to unsubstantiated advice being provided to users. In addition, preferences are an essential object in many contexts, such as decision-making, machine learning, recommendation systems, social choice theory, and various sub-fields of artificial intelligence. In this context, the challenge is to build learning algorithms that are both efficient (from a computational point of view) while keeping humans in the loop to integrate and represent as faithfully as possible their expertise and their skills Knowledge.

  The basic idea of the multi-criteria decision support methodology is that, given a decision problem, we collect preferential information from the decision-maker to build

an evaluation model that must reflect the point of view. (the value system) of the decision-maker and help him solve his decision problem. In other words, my research is interested in implementing algorithms for the automatic learning of preferences based on reference examples (a training set). Several models are studied: sorting, classification and point of reference models. To answer the question, different tools and methods are used for the formulation of preference learning algorithms: mathematical programming and logical formulations (SAT / MAxSAT).

  – **Theses concerned**: Ali Tlili (2022), Pegdwendé Stéphane Minoungou (2022), Jinyan Liu (2016)

- **Design of adaptive dialogue protocols**: decision support is an interaction between at least two agents. Setting up an automatic system to support this interaction raises several questions: how to model the system's reasoning to allow "efficient" interaction with a user; how to make a formal link between the generation of the explanation and the improvement of the learning process. Indeed, faced with an explanation, a user can provide new information, invalidate old information, etc. These reactions strongly contribute to feeding other phases of the decision support process, such as the learning phase of the preference model. How to adapt classic preference learning algorithms to manage inconsistent user feedback (inconsistency, erroneous information, etc.) while automatically adjusting the model to the information provided by the user?

  In this context, my research aims to provide a formal language to represent such an interaction, explain it, communicate its results, and convince the user that what is happening is both theoretically sound and operationally reasonable. To do this, we propose to build and formalize an interaction protocol, which specifies the rules and conditions under which we can have a "coherent" interaction in a decision support context where the initiative is sometimes left to the user (e.g. ask for an explanation). We will rely on dialectical management and dialogue systems resulting from work in multi-agent systems and argumentation theory.

  – **Theses concerned**: Manuel Amoussou (in progress).

Finally, through the previous axes, our ambition is to obtain solid theoretical frameworks. Beyond this, we wish to prove the utility and the applicability of the theoretical propositions through real situations. The objective is to offer algorithmic solutions to real-world problems by combining multicriteria decision support tools and artificial intelligence.

- **Theses concerned**: Ali Tlili (2022), Mathieu Lerouge (in progress), Manel Mammar (2015), Massinissa Mammeri (2017)

Supervision

Thesis in progress

- Dao Thauvin.Explanatory dialogue for the interpretation of visual scenes (Funded AID-ONERA). Co-supervised with 15% with Stéphane Herbin (ONERA) and Céline Hudelot (MICS, CentraleSupélec). (Start November 2022).

- Mathieu Lerouge. Designing explanation schemes for recommendations stemming from Optimization Systems: application to scheduling problems for facility management (MICS, CentraleSupélec- Decision Brain). Funding PSPC AIDA Project. Co-supervision 30% with Vincent Mousseau (MICS-CentraleSupélec), Céline Gicquel (LISN, Université Paris Saclay) (start December 2020).

- Manuel Amoussou. Interactive explanations in Multi-criteria decision aiding: handling inconsistencies and levels of explanation. (MICS, CentraleSupélec). Funding PSPC AIDA Project. Co-supervision 50% with Vincent Mousseau (MICS-CentraleSupélec) (start May 2020). **Publications:** [34].

Defended Thesis

- Ali Tlili (15/06/2022). Multicriteria Portfolio Management Optimization (MICS, Centrale-Supélec - Dassault Systèmes). Funding Dassault Systèmes. Co-supervision à 50% with Vincent Mousseau (MICS, CentraleSupélec), and Khaled Oumeima (Dassault Systèmes[2]).

    - **Publications:** [3], [4], [38].

    - **Job**: Operational Research Technology Specialist (Dassault Systèmes)

- Pegdwendé Stéphane Minoungou (13/05/2022). Learning an MR-Sort model from non monotone data (MICS, Centalesupélec -IBM Zurich). Funding IBM. Co-supervision 50% with Vincent Mousseau (MICS, CentraleSupélec) and Paolo Scoton (IBM Zurich).

    - **Publications:** [2], [33].

    - **Job**: Research Engineer, since 2022 (Anse Technology).

- Ismaïl Baaj (27/01/2022). Explainability of possibilistic and fuzzy rule-based systems. (LIP6, Sorbonne Université- CEA List - MICS, CentraleSupélec). Funding CEA. Co-supervision 30% with Nicolas Maudet (LIP6, Sorbonne Université) and Jean-Philippe Poli (CEA List[3]).

    - **Publications:** [14], [16], [35].

    - **Job**: Post-Doc Telcome SudParis.

- Khaled Belahcène (05/12/2018). A contribution to accountable decision aiding : explanations for the aggregation of preferences (LGI, CentraleSupélec - LIP6, Sorbonne Université). Doctoral School INTERFACES research grant funding. Co-supervision (25%) with Vincent Mousseau (LGI, CentraleSupélec), Nicolas Maudet (Sorbonne Université) and Christophe Labreuche (Thales Research and Technology).

    -

    - **Publications**: [4], [5], [7], [9], [17], [18], [19], [34], [36], [37], [39].

    - **Job**: Assistant Professor since 2019, Heudiasyc[4], UTC.

- Massinissa Mammeri (28/11/2017). Decision aiding methodology for developing the contractual strategy of complex oil and gas projects (LGI, CentraleSupélec - Total). Funding Total. Co-supervision 50% with Franck Marle (LGI, CentraleSupélec).

    - **Publications**: [22]

    - **Job**: Business Intelligence Consultant since 2017 (SYSTRA).

- Karim El Mernissi (13/12/2017). Generation of explanations in rule-based systems (LIP6-UPMC, LGI-CentraleSupélec, IBM). Funding IBM. Université Pierre et Marie Curie. Co-supervision 50% with Nicolas Maudet (LIP6, UPMC) and Pierre Feillet (IBM)

    - **Publications**: [20]

    - **Job**: Data Scientist since 2019 (Orange, paris).

- Jinyan Liu (09/03/2016). Elicitation of preferences for a model based on reference points (LGI, Ecole Centrale Paris). Funding CSC scholarship. Co-supervision 50% with Vincent Mousseau (LGI, Ecole Centrale Paris).

    - **Publications**: [8], [25], [40].

    - **Job**: Tech Lead Data Scientist since 2019 (Faurecia, Paris).

---

[2]https://www.3ds.com
[3]http://www-list.cea.fr/en/
[4]https://www.hds.utc.fr/en.html

- Manel Maamar (07/12/2015). Multi-criteria modeling and optimization with anticipation of a Leads marketplace (LGI, Ecole Centrale Paris). Funding Place des Leads. Co-supervision 50% with Vincent Mousseau (LGI, Ecole Centrale Paris) and Alexandre Aubry (Place des Leads).

    - **Publications**: [24]
    - **Job**: Machine Learning Consultant since 2019 (Groupe Pact Novation, Paris).

Master Thesis

- Nathan Rougier. Artificial Intelligence methods for prediction and management of patient flows in hospital departments (MICS, CentraleSupélec). M2 (third year engineering). In collaboration with Gianluca Quercini (LISN, Université Paris Saclay). Supervision 70%. CentraleSupélec, 2021-2022. DataIA Funding.

- Antonin Billet, "Evaluation of a conceptual model of Fake News". May- july 2022 at St-Cyr Coëtquidan (M1). (33% with Nicolas Belloir, Saint-Cyr, IRISA and Oscar Pastor, PROSS, Universidad Politécnica de Valencia, Spain).

- Evan Epivent, "Towards an XAI approach based on a conceptual model of Fake News". Stage de M1 à St-Cyr Coëtquidan. June- September 2022 (M1). (33% with Nicolas Belloir, Saint-Cyr, IRISA and Oscar Pastor, PROSS, Universidad Politécnica de Valencia, Spain).

- Emilien Frugier. "Conceptual Modelling of Fake News". 2021-2022. Double Diploma St-Cyr Coëtquidan-CentraleSupélec (M2). (33% with Nicolas Belloir, Saint-Cyr, IRISA and Oscar Pastor, PROSS, Universidad Politécnica de Valencia, Spain).

- Antonin Duval. Deep reinforcement learning in the multi-agent framework in simulations (Thales Research & Technology). Msc IA[5]. Supervision 100%. CentraleSupélec, 2019-2020.

- Sanae Chouhani. Optimization of train movement in technicenter (SNCF). Master 2 OSIL. Supervision 100% CentraleSupélec, 2017-2018.

- Rihab Brahim. Improvement of industrial planning processes (LVMH). Master 2OSIL. Co-supervision (30%) with Yves Dallery. 2016-2017.

- Léonel de la Bretesche. Optimization method from an outsourced warehouse Application to the case of the Amazon-SMOBY warehouse (AMAZON). Master 2 OSIL. Supervision 100%. École Centrale Paris, 2014-2015.

- Massinissa Mammeri. Lead forecasting problem for a marketplace (Place des Leads). Master 2 MODO (Modélisation, Optimisation, Décision et Organisation). Co-supervision (25%) avec Denis Bouyssou (Université paris dauphine), Vincent Mousseau (ECP), Alexandre Aubry (Place des Leads). Université Paris-Dauphine. 2013-2014.

- Lisa JUNGE. Hybridization and electrification of CLAAS tractors: potentials and economic prospects, (CLAAS Tractor SAS). Master 2 OSIL. Supervision 100%. Ecole Centrale Paris, 2012-2013.

- Liu Jinyan. Inference of a multi-criteria multi-decision maker ranking: a method based on reference points. Research internship. Master 2 OSIL. Co-supervision (50%) with Vincent Mouseau. Ecole Centrale Paris, 2011-2012.

- Bian Yuan. Multiple criteria models for competence-based project staffing. Research internship. Master 2 OSIL (Optimisation des Systèmes Industriels et Logistiques), co-supervision (50%) with Vincent Mousseau. Ecole Centrale Paris, 2011-2012

---

[5]https://www.centralesupelec.fr/fr/msc-artificial-intelligence

	Number
Theses in progress	03
Defended Theses	08
Master2 Theses	10
Master1 Theses	10

Table 1: Supervisions summary

DISSEMINATION AND RESPONSIBILITIES

Contracts

- Funding of an M2 internship by the "M2 2022 internship call" of DataIA[6]. Subject: Artificial Intelligence methods for the prediction and management of patient flows in hospital services. In collaboration with Gianluca Quercini (LISN, Unviersité Paris Saclay).

- Scientific coordinator of WP-F (Generation and representation of explanations by the AIDA System) of the PSPC AIDA (AI for Digital Automation) project carried by IBM (MICS budget - 320k€). Start January 2020 (48 months).

- Coordination of a proposal in response to the "Expression of Interest - IBM Research Collaborations" through DATAIA[7]. This proposal resulted in the funding (120k€) of a CIFRE thesis which began in March 2019 in co-supervision with Vincent Mousseau (MICS, CentraleSupélec) and Paolo Scoton (IBM Zurich).

Prize and Distinction

- RCIS 2022 Best Forum Paper / Poster Award

- Doctoral and Research Supervision Bonus (2020-2024)

- Doctoral and Research Supervision Bonus (2015-2019)

Member of a Jury thesis

- Thesis of Fabien de Lacroix. Title: Dialogue to decide. Proactive expert recommendation and fair multi-agent decision making. (Université Lille 1, 2015).

- Thesis of Olivier Sobrie. Title: Learning preferences with multiple-criteria models (Université de Mons, 2016).

- Thesis of Tasneem Bani-Mustapha. Title: multi-hazards risk aggregation considering trustworthiness of the assessment (LGI, CentraleSupélec, 2019).

Participation in committees

- **Guest Editor** pour EURO Journal on Decision Processes (EJDP), Special issue: Supporting and Explaining Decision Processes by means of Argumentation 2018.

- **Reviewer for International Journals** : Journal of Autonomous Agents and Multi-Agent Systems, Multi-Criteria Decision Analysis (JMCDA), Annals of Operations Research, European Journal of Operation Research (EJOR), Argument and Computation, Operational Research - An International Journal (ORIJ), The International Journal of Management Science (OMEGA), Transaction on Fuzzy Systems.

---

[6]https://www.dataia.eu/appel-projets/appel-stages
[7]https://dataia.eu

- **PC international conferences and workshops** : AAAI (2021, 2020, 2019), AAMAS (2019), IJCAI (2022, 2021 (SPC), 2020, 2019, 2018), KR (2018), ECAI (2020), IPMU (2012), DA2PL[8] (2020, 2018, 2016, 2012).
- **PC national conferences and workshops** : JFSMA (2022, 2021, 2020), RJCIA (2018, 2016, 2017), MFI (2013).

Participation, Presentations in conferences and seminars

- Wassila Ouerdane. Title: Generation of Textual Explanations in XAI: the Case of Semantic Annotation. Explicability and symbolic reasoning in AI" seminar for the D2K[9] working group, from Data to Knowledge, resumes its meetings. 23 November 2021
- Wassila Ouerdane. Title: The challenges of "intelligent" decision support: from preference learning to explaining recommendations. Journée "Philosophie des sciences et Intelligence Artificielle[10]" (PS & IA 2020). 06 Feverier 2020.
- Wassila Ouerdane. Title: A Dialogue Game for Recommendation with Adaptive Preference Models. MICS Seminar. 24 June 2019.
- Wassila Ouerdane et Vincent Mousseau. Title: Interactive Recommendation and Explanation for Multiple Criteria Decision Analysis. Séminaire IRT SystemX[11]. 11 april 2018.
- Wassila Ouerdane. Title: Justified decisions are better than simple ones: explaining preferences using even swap sequences. In 26th European Conference on Operational Research. Rome, Italie. 1-4 July, 2013. Join work with Christophe Labreuche, Nicolas Maudet and Vincent Mousseau.

Working Groups

- Member of the National French Research Group in IA 'Explainability' working group (https://gt-explication.gitlab.io/)
- Member of of the National French Research Group in IA (https://www.gdria.fr).

TEACHING

Since my recruitment as a lecturer (assistant professor), I had taught or taught at all university levels (Bachelor, Master) in the IT department at CentraleSupélec (when I arrived, École Centrale Paris). I am also involved in the Master of Science Artificial Intelligence [12] of CentraleSupélec. The summary of the teaching hours is presented in the Table3. I also supervise a number of end studies internship, gap year and group projects.

The number of hours mentioned in this table count the equivalent hours of tutorials performed, generally distributed in lessons, tutorials and for certain courses in practical work and project monitoring. I would like to point out that this service was impacted by three maternity leaves: from January 17, 2011 to May 7, 2011; from October 17, 2014 to February 8, 2015 and from September 19, 2020 to March 18, 2021.

List of Current Courses and activities–2021/2022

- Information retrieval and processing of big data –112 students. Co-leader with Céline Hudelot (MICS, CentraleSupélec)

---

[8] From Multiple Criteria Decision Aid to Preference Learning - https://event.unitn.it/da2pl2020/#home
[9] https://digicosme.cnrs.fr/event/groupe-de-travail-de-la-donnee-a-la-connaissance/
[10] https://afia.asso.fr/psia-2020/
[11] https://www.youtube.com/watch?v=it50bttu4P8
[12] https://www.centralesupelec.fr/fr/msc-artificial-intelligence

Period	Bachelor Level	Master Level	Total
2010-2011	85	36	121
2011-2012	67	150	217
2012-2013	130	150	280
2013-2014	67	150	217
2014-2015	85	33	118
2015-2016	120	158	278
2016-2017	125	126	250
2017-2018	112	135	247
2018-2019	112	135	247
2019-2020	200	50	250
2020-2021	78	32	110

Table 2: Summary Teaching hours

- Multi-agent system: architectures and reasoning –Master level, shared with the MSc Artificial Intelligence, 55 students. Course leader

- Explainability of AI Systems - Master level, 60 students. Co-leader with Jean-Philippe Poli (CEA List)

- SAFRAN AI Training: "Multi-agent Systems" (16 participants) 2021 and 2022.

- DGA AI Training: "Autonomous Agents and Decision Aiding" (10 participants) 2022.

Publications

Wassila OUERDANE

November 2022

Articles under submission

- Khaled Belahcène, Vincent Mousseau, Wassila Ouerdane, Marc Pirlot, and Olivier Sobrie. Multiple Criteria Sorting: a model-oriented survey. Submitted to 4OR (October 2022)

- Manuel Amoussou, Khaled Belahcène, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Computing explanations for a multicriteria additive value based model. Submitted to EJOR (September 2022).

- Mathieu Lerouge, Céline Gicquel, Vincent Mousseaun and Wassila Ouerdane. Explaining solutions stemming from optimization systems solving the Workforce Scheduling and Routing Problem to their end-users. Submitted to EJOR (July 2022)

Articles published in international peer-reviewed journals

[1] Khaled Belahcène, Vincent Mousseau, Wassila Ouerdane, Marc Pirlot, Olivier Sobrie, Ranking with Multiple Reference Points: Efficient SAT-based learning procedures, Computers & Operations Research, Volume 150, 2023.

[2] Pegdwendé Minoungou, Vincent Mousseau, Wassila Ouerdane, Paolo Scotton. A MIP-based approach to learn MR-Sort models with single-peaked preferences. Annals of Operations Research, Springer Verlag, 2022. https://doi.org/10.1007/s10479-022-05007-5

[3] Ali Tlili, Oumaima Khaled, Vincent Mousseau, and Wassila Ouerdane. Interactive portfolio selection involving multicriteria sorting models. Ann Oper Res (2022). https://doi.org/10.1007/s10479-022-04877-z

[4] Ali Tlili, Khaled Belahcène, Oumaima Khaled, Vincent Mousseau, Wassila Ouerdane: Learning non-compensatory sorting models using efficient SAT/MaxSAT formulations. European Journal of Operational Research 298(3): 979-1006 (2022)

[5] Alexandru-Liviu Olteanu, Khaled Belahcène, Vincent Mousseau, Wassila Ouerdane, Antoine Rolland, Jun Zheng: Preference elicitation for a ranking method based on multiple reference profiles. 4OR 20(1): 63-84 (2022) .

[6] Anthony Hunter, Nicolas Maudet, Francesca Toni, Wassila Ouerdane. Foreword to the Special Issue on supporting and explaining decision processes by means of argumentation. EURO journal on decision processes, Volume 6, Issue 3–4, pp 235–236, 2018.

[7] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. An efficient SAT formulation for learning multiple criteria non-compensatory sorting rules from examples. Computers and Operations Research, Elsevier, Volume 97, pp 58-71, 2018.

[8] Valentina Ferretti, Liu Jinyan, Vincent Mousseau, Wassila Ouerdane. Reference-based ranking procedure for environmental decision making: Insights from an ex-post analysis. Environmental Modelling and Software, Elsevier, Volume 99, pp.11-24. 2018.

[9] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. Explaining robust additive utility models by sequences of preference swaps. Theory and Decision, Springer Verlag, Volume 82, Issue 2, pp 151-183, 2017.

[10] Wassila Ouerdane, Yannis Dimopoulos, Konstantinos Liapis, Pavlos Moraitis. Towards automating Decision Aiding through Argumentation. Journal of Multicriteria Decision Analysis, Volume 18, pp 289-309, 2011.

[11] Wassila Ouerdane. Multiple Criteria Decision Aiding: a Dialectical Perspective. 4OR: A Quarterly Journal of Operations Research, Springer Verlag, Volume 9, Issue 4, pp 429–432, 2011.

Articles published in international conferences with peer review

[12] Nicolas Belloir, Wassila Ouerdane, and Oscar Pastor. Characterizing Fake News: A Conceptual Modeling-based Approach. In proceedings of the 41ST internatinal conference on Conceptual Modeling (ER) 2022. (to appear).

[13] Nicolas Belloir, Wassila Ouerdane, Oscar Pastor, Emilien Frugier, Louis-Antoine de Barmon, A Conceptual Characterization of Fake News: A Positioning Paper. In: Guizzardi, R., Ralyté, J., Franch, X. (eds) Research Challenges in Information Science. RCIS 2022. Lecture Notes in Business Information Processing, vol 446.pp 662–669. Springer, Cham. 2022. (*RCIS 2022 Best Forum Paper / Poster Award*).

[14] Ismaïl Baaj, Jean-Philippe Poli, Wassila Ouerdane, Nicolas Maudet. Representation of Explanations of Possibilistic Inference Decisions. ECSQARU 2021: European Conference on Symbolic and Quantitative Approaches with Uncertainty, Sep 2021, Prague, Czech Republic. pp.513-527.

[15] Jean-Philippe Poli, Wassila Ouerdane, Regis Pierrard. Generation of Textual Explanations in XAI: the Case of Semantic Annotation. 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jul 2021, Luxembourg, Luxembourg. pp.9494589

[16] Ismaïl Baaj, Jean-Philippe Poli, Wassila Ouerdane, Nicolas Maudet. Min-max inference for Possibilistic Rule-Based System. 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jul 2021, Luxembourg, Luxembourg. pp.9494506.

[17] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. Comparing options with argument schemes powered by cancellation. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19), Macao,China. pp 1537-1543, 2019.

[18] Khaled Belahcène, Yann Chevaleyre, Nicolas Maudet, Christophe Labreuche, Vincent Mousseau, and Wassila Ouerdane. Accountable Approval Sorting. Proceedings of 27th International Joint Conference on Artificial Intelligence and 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018). Stockholm, Sweden. pp 70-76, 2018.

[19] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau and Wassila Ouerdane. A Model for Accountable Ordinal Sorting. In proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-2017), Melbourne, Australia. pp 814-820, 2017.

[20] Karim El Mernissi, Pierre Feillet, Nicolas Maudet, Wassila Ouerdane. Introducing Causality in Business Rule-Based Decisions. In proceedings of the 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2017), Arras, France. Springer, Advances in Artificial Intelligence: From Theory to Practice: pp.433-439, 2017.

[21] Mathieu Dernis, Wassila Ouerdane, Ludovic-Alexandre Vidal, Pascal Da Costa, Franck Marle. Assessment of Sustainable Strategies based on DMM Approach and Value Creation. In 19th International Dependency and Structure Modelling Conference (DSM), Helsinki, Finland. Understand, Innovate, and Manage your Complex System! 2017.

[22] Massinissa Mammeri, Franck Marle, Wassila Ouerdane. An assistance to identification and estimation of contractual strategy alternatives in oil and gas upstream development projects. In 19th International Dependency and Structure Modelling Conference (DSM), Helsinki, Finland. 2017, Understand, Innovate, and Manage your Complex System. 2017.

[23] Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane, Simon Parsons. A dialogue game for recommendation with adaptive preference models. In proceeding of the 14th International Conference on Autonomous Agents and Multiagent systems. Istanbul, Turkey. pp.959-967. 2015.

[24] Manel Mammar, Vincent Mousseau, Wassila Ouerdane, Alexandre Aubry. Internet Prospect's flow forecasting for a multi-period optimization model of offer/Demand assignment problem. International Conference on computers and Industrial Engineering (CIE45), Oct 2015, Metz, France.

[25] Jinyan Liu, Vincent Mousseau, Wassila Ouerdane. Preference Elicitation from Inconsistent Pairwise Comparisons for Multi-criteria Ranking with Multiple Reference Points. In proceedings of the 14th International Conference on informatics and Semiotics in Organisations. Web of thingd, People and Information Systems (ICISO), Stockholm, Sweden. pp 120-130, 2013.

[26] Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane. Justifying Dominating Options when Preferential Information is Incomplete. Proceedings of the 20th European Conference on Artificial Intelligence (ECAI'12), Montpellier, France. IOS Press, 242, pp.486-491, Frontiers in Artificial Intelligence and Applications. 2012.

[27] Myriam Merad, Wassila Ouerdane, Nicolas Dechy. Expertise and decision-aiding in safety and environment domains: what are the risks?. BERENGUER, C.; GRALL, A. ; GUEDES SOARES, C. Proceedings of The annual European Safety and Reliability (ESREL) conference. Troyes, France. CRC Press. London, pp.2317-2323, 2011.

[28] Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane. Minimal and Complete Explanations for Critical Multi-attribute Decisions. In Proceedings of the

2nd International Conference on Algorithmic Decision Theory (ADT'2011), Piscataway New Jersey, United States. Springer, Lecture Notes in Computer Science. pp.121-134, 2011.

[29] Wassila Ouerdane, Nicolas Maudet, Alexis Tsoukiàs. Dealing with the dynamics of proof-standard in argumentation-based decision aiding. Proceedings of the 5th European Starting AI Researcher Symposium (STAIRS'10). co-located with ECAI 2010, Lisbon, Portugal. IOS Press, pp.225-237. 2010.

[30] Wassila Ouerdane, Nicolas Maudet and Alexis Tsoukiàs. Dealing with the dynamics of proof-standard in argumentation-based decision aiding. Proceedings of 19th European Conference on Artificial Intelligence (ECAI'10).Frontiers in Artificial Intelligence and Applications, IOS Press. Lisbon, Portugal. pp. 999-1000, 2010.

[31] Wassila Ouerdane, Nicolas Maudet, Alexis Tsoukiàs. Argument Schemes and Critical Questions for Decision Aiding Process. Proceedings of the 2nd international conference on Computational Models of Argument (COMMA2008), Toulouse, France. pp. 285-296, 2008

[32] Wassila Ouerdane, Nicolas Maudet, Alexis Tsoukias. Arguing over actions that involve multiple criteria: A critical review. In Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'07), Hammamet, Tunisia. pp.308–319, 2007.

Articles published in international workshops with peer review

[33] Pegdwendé Minoungou, Vincent Mousseau, Wassila Ouerdane, and Paolo Scotton. Learning an MR-Sort model from data with latent criteria preference direction. In the 5th workshop from multiple criteria Decision aid to Preference Learning (DA2PL), 5-6 November, 2020. University of Trento, Trento - Italy.

[34] Manuel Amoussou, Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau and Wassila Ouerdane. Explaining Robust Additive Decision Models: Generation of Mixed Preference-Swaps by Using MILP. In the 5th workshop from multiple criteria Decision aid to Preference Learning (DA2PL), 5-6 November, 2020. University of Trento, Trento - Italy.

[35] Ismaïl Baaj, Jean-Philippe Poli and Wassila Ouerdane. Some Insights Towards a Unified Semantic Representation of Explanation for eXplainable Artificial Intelligence (XAI). Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI). Association for Computational Linguistics. Tokyo, Japan. pp 14-19, 2019.

[36] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau and Wassila Ouerdane. Challenges in Interactive Explanation and Recommendation for Decision Support. In The international Workshop on Dialogue, Explanation and Argumentation in Human-Agent Interaction (DEXAHAI [1]) Southampton UK. 2018.

[37] Khaled Belahcène, Vincent Mousseau, Wassila Ouerdane, Marc Pirlot and Olivier Sobrie. Ranking with Multiple Points: Efficient Elicitation and Learning Procedures. In the 4th workshop from multiple criteria Decision aid to Preference Learning (DA2PL), 2018. Poznan, Pologne.

[38] Khaled Belahcène, Oumaima Khaled, Vincent Mousseau, Wassila Ouerdane and Ali Tlili. A new efficient SAT formulation for learning NCS models: numerical results. In the 4th workshop from multiple criteria Decision aid to Preference Learning (DA2PL), 2018. Poznan, Pologne.

[39] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau and Wassila ouerdane. Accountable classifications without frontiers. In the 3rd workshop, euro mini group, from multiple criteria Decision aid to Preference Learning (DA2PL), 2016, Paderborn, Germany.

[40] Jinyan Liu, Wassila Ouerdane, Vincent Mousseau. A Methaheuristic approach for preference Learning in multi criteria ranking based on reference points. In the 2nd workshop from multiple criteria Decision aid to Preference Learning (DA2PL), Nov 2014, Chatenay Malabry, France.

Articles in international conferences or workshops (extended abstract)

[41] Khaled Belahcène, Vincent Mousseau, Wassila Ouerdane, Marc Pirlot and Olivier Sobrie. Ranking with Multiple Points: Efficient Elicitation and Learning Procedures. In the 25th International Conference on Multiple Criteria Decision-Making (MCDM), Istanbul, Turkie, 2019.

[42] Manel Mammar, Vincent Mousseau et Wassila Ouerdane. Multicriteria Modeling and Optimization of a market place of leads. In 22nd International Conference on Multiple Criteria Decision Making. Malaga (Spain) 17-21 juin 2013.

[43] Jinyan Liu, Vincent Mousseau and Wassila Ouerdane. Titre: Robust Elicitation of a Qualitative Ranking Model using Inconsistent Data. Dans 22nd International Conference on Multiple Criteria Decision Making . Malaga (Spain). 17-21 juin 2013.

---

[1]https://sites.google.com/view/dexahai-18/home

[44] Manel Mammar, Vincent Mousseau et Wassila Ouerdane. Titre: Modélisation et optimisation multicritère d'une place de marché de Leads (Adéquation offre/demande). Dans 77th meeting of the European working group on multicriteria decision aiding (MCDA'77). Rouen, 2013.

[45] Jinyan Liu, Vincent Mousseau and Wassila Ouerdane. Titre: Preference Elicitation for Multi-Criteria Ranking with Multiple Reference Points. Dans 77th meeting of the European working group on multicriteria decision aiding (MCDA'77). Rouen, 2013.

Articles published in National conferences or workshops with peer review

[46] Manuel Amoussou, Khaled Belahcène, Nicolas Maudet, Vincent Mousseau and Wassila Ouerdane. Des explications par étapes pour le modèle additif. Journées d'Intelligence Artificielle Fondamentale (JIAF), 2022, Saint-Étienne, France (https://hal.archives-ouvertes.fr/hal-03781382/document).

[47] Mathieur Lerouge, Céline Giquel, Vincent Moussea, and Wassila Ouerdane. "Designing methods for explaning solutions stemming from optimization systems, application to the workforce and scheduling routine", at the annual congress in Operations Research and Decision Support ROADEF 2022, organized by the French association ROADEF, on February 23rd to 25th 2022, in Lyon.

[48] Jean-Philippe Poli, Wassila Ouerdane, et Régis Pierrard. Génération d'explications textuelles en XAI : le cas de l'annotation sémantique. Dans LFA 2021 Rencontres Francophones sur la Logique Floue et ses Applications, October 2021, Paris, France.

[49] Ismail Baaj, Jean-Philippe Poli, Wassila Ouerdane and Nicolas Maudet. . Inférence min-max pour un système à base de règles possibilistes. Dans LFA 2021 Rencontres Francophones sur la Logique Floue et ses Applications, October 2021, Paris, France.

[50] Khaled Belahcène, Yann Chevaleyre, Nicolas Maudet, Christophe Labreuche, Vincent Mousseau and Wassila Ouerdane. Accountable Approval Sorting. Dans le 20me congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF'2019). Havre, France.

[51] Khaled Belahcène, Oumaima Khaled, Vincent Mousseau, Wassila Ouerdane and Ali Tlili. A new efficient SAT formulation for learning NCS models: numerical results. Dans le 20me congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF'2019). Havre, France.

[52] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. Une formulation SAT pour l'apprentissage de modèles de classement multicritères noncompensatoires. 11e Journées d'Intelligence Artificielle Fondamentale, Jul 2017, Caen, France.

[53] Mathieu Dernis, Ludovic-Alexandre Vidal, Franck Marle, Wassila Ouerdane, Pascal Da Costa. Aide à la sélection de stratégies pour apporter des valeurs durables à des pays hôtes en contexte pétrolier. Congrès International de Génie Industriel CIGI, May 2017, Compiègne, France. 2017.

Book Chapter

[54] Wassila Ouerdane et al. Recherches en IA explicable au MICS: Modèles gaussiens, modèles génératifs et raisonnement pourl'explicabilité. Association française pour l'Intelligence Artificielle. 2022. IA & Explicabilité. Bulletin de l'AFIA, 116, 62.

[55] Wassila Ouerdane, Nicolas Maudet, Alexis Tsoukias. Argumentation Theory and Decision Aiding. J. Figueira, S. Greco, and M. Ehrgott. Trends in Multiple Criteria Decision Analysis, 142 (1), pp.177-208, 2010, International Series in Operations Research and Management Science.

PhD Thesis

[56] Wassila Ouerdane. Multiple Criteria Decision Aiding : a Dialectical Perspective. Thèse de Doctrorat. Université Paris Dauphine - Paris IX, Deccember, 2009.

	Number	Acronym/Name
International Journal	11	EJOR, 40R, EJDP, COR, Environmental Modelling & Software, Theory and Decision, JMCDA, Annals of OR
International Conferences	21	IJCAI 2019, 2018, 2017 (A*), AAMAS 2015 (A*), ER 2022 (A), ECAI 2012, 2010 (A), RCIS 2022 (B), FuzzyIEEE 2021 (B), IEA/AIE'17 (C), ECSQARU 2021, 2007 (C), ADT 2011, COMMA 2008, DSM 2017, ICISO 2013, ESREL 2011, STAIRS 2010
International Workshops	08	DA2PL 2020, 2018, 2016, NL4XAI 2019, DEXAHAI 2018,
National Workshps	06	LFA2021, ROADEF 2019, 2022, JIAF 2017,2022, CIGI 20217
Book chapter	02	Bulletin AFIA, Trends in Multiple Criteria Decision Analysis

Table B.1: Publications Summary

Selection of Articles

Please find below a selection of articles organized by chapters and sorted in descending years.

C.1 Selection of articles related to Chapter 3

- Ali Tlili, Khaled Belahcène, Oumaima Khaled, Vincent Mousseau, Wassila Ouerdane: Learning non-compensatory sorting models using efficient SAT/MaxSAT formulations. European Journal of Operational Research 298(3): 979-1006 (2022)

- Alexandru-Liviu Olteanu, Khaled Belahcène, Vincent Mousseau, Wassila Ouerdane, Antoine Rolland, Jun Zheng: Preference elicitation for a ranking method based on multiple reference profiles. 4OR 20(1): 63-84 (2022).

- Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. An efficient SAT formulation for learning multiple criteria noncompensatory sorting rules from examples. Computers and Operations Research, Elsevier, Volume 97, pp 58-71, 2018.

- Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. Explaining robust additive utility models by sequences of preference swaps. Theory and Decision, Springer Verlag, Volume 82, Issue 2, pp 151-183, 2017.

Decision Support

Learning non-compensatory sorting models using efficient SAT/MaxSAT formulations

Ali Tlili [a], Khaled Belahcène [b], Oumaima Khaled [a], Vincent Mousseau [a,*], Wassila Ouerdane [a]

[a] *MICS, CentraleSupélec, Université Paris-Saclay, Gif-Sur-Yvette, France*
[b] *Université de Technologie de Compiègne, CNRS, Heudiasyc, France*

A B S T R A C T

The Non-Compensatory Sorting model aims at assigning alternatives evaluated on multiple criteria to one of the predefined ordered categories. Computing parameters of the Non-Compensatory Sorting model compatible to a set of reference assignments is computationally demanding. To overcome this problem, two formulations based on Boolean satisfiability have recently been proposed to learn the parameters of the Non-Compensatory Sorting model from perfect preference information, i.e. when the set of reference assignments can be completely represented in the model. In this paper, two popular variants of the Non-Compensatory Sorting model are considered, the Non-Compensatory Sorting model with a unique profile and the Non-Compensatory Sorting model with a unique set of sufficient coalitions. For each variant, we start by extending the formulation based on a separation principle to the multiple category case. More-over, we extend the two formulations to handle inconsistency in the preference information using the Maximum satisfiability problem language. A computational study is proposed to compare the efficiency of both formulations to learn the two Non-Compensatory Sorting models (with a unique profile and with a unique set of sufficient coalitions) from noiseless and noisy preference information.

© 2021 Elsevier B.V. All rights reserved.

Introduction

Multiple Criteria Decision Analysis (MCDA) aims at developing decision-support models explicitly based on the construction of a set of criteria reflecting the relevant aspects of the decision-making problem. These n criteria ($\mathcal{N} = \{1, 2, \ldots, n\}$ with $n \geq 2$) evaluate a set of alternatives $A = \{a, b, c, \ldots\}$ under consideration with respect to different viewpoints. The MCDA literature considers different problem statements to formulate real-world decision problems; Roy (1996) distinguishes three problem statements: choice, sorting and ranking. As opposed to choice and ranking problem formulations which are comparative in nature, sorting formulates the decision problem in terms of the assignment of alternatives to one of the predefined ordered categories $C^1, C^2, \ldots C^p$, where C^1 (C^p, resp.) is the worst (the best, resp.) category. The assignment of an alternative to the appropriate category relies on its intrinsic value, and not on its comparison with other alternatives

In this paper, we are interested in a specific sorting procedure: the Non-Compensatory Sorting (NCS) model (Bouyssou & Marchant, 2007a; 2007b), which corresponds to a generalization

and formal description of the Electre Tri procedure (Figueira, Greco, Roy, & Słowiński, 2010). One of its specificity is to account for the alternative evaluations in an ordinal perspective avoiding compensation and enables to deal meaningfully with qualitative data.

We consider a decision aiding process in which two participants are involved to model a sorting problem using NCS: a decision maker (DM) looking for a recommendation and an analyst to support the DM in the search for this recommendation. Thus, the role of the analyst is to interact with the DM in order to help her to elaborate her preferences which are generally not fully predefined at the beginning of the decision process. The DM expresses preferences from which a specific NCS model is inferred. More specifically, the information supplied by the DM in order to specify the NCS sorting model are assignment examples (alternatives that should be assigned to a category). It should be highlighted that the construction of the learning set and the NCS model often results from a sequence of interactions between the DM and the analyst rather than in a one-step interaction.

In this perspective, the inverse Non-Compensatory Sorting problem (Inv-NCS, see Section 3.1) takes as input a set of assignment examples, and computes (whenever it exists) an NCS sorting model which is consistent with this preference information. In other words, Inv-NCS learns the NCS parameters that perfectly match a set of desired outputs (assignment examples). Solving

---

Inv-NCS problem is computationally difficult and have been proved to be NP-hard (Belahcene et al., 2018b). Mixed-integer linear formulations (Leroy, Mousseau, & Pirlot, 2011; Zheng, Metchebon Takougang, Mousseau, & Pirlot, 2014) and heuristic resolution approaches (Sobrie, Mousseau, & Pirlot, 2015; 2019) have been proposed for Inv-NCS.

Recently, Belahcene, Labreuche, Maudet, Mousseau, & Ouerdane (2018a) proposed a SAT formulation of this problem which proves to be computationally more efficient than previous approaches. In this paper, we report a second SAT formulation for Inv-NCS (Belahcene et al., 2018b) described in the context of two categories. We extend this second formulation to the multiple category case and perform numerical tests to compare the performance of these two SAT formulations. Nevertheless, in the SAT problem, it is assumed that the set of assignment examples is fully compatible with NCS. Therefore, we are interested in extending both SAT formulations to handle inconsistency in preference information which is often the case in real-world decision problems.

Indeed, in actual case studies, preference expressed by DMs are often inconsistent, due to the multiplicity of DMs, the fact that their preferences are not necessarily predefined and can evolve during the elicitation process. Handling inconsistencies when considering a set of preference statements on a set of multicriteria alternatives has been already tackled in the literature through mathematical programming (see, e.g. Mousseau, Dias, Figueira, Gomes, & Clímaco, 2003) or the analysis of reciprocal preference relations (see, e.g. Herrera-Viedma, Herrera, Chiclana, & Luque, 2004). In this work, we consider handling inconsistency with MaxSAT language, where a SAT formulation is complemented with an implicit objective function, so that the number of satisfied clauses is maximal, allowing to best satisfy an unsatisfiable instance and consequently to best restore the assignment examples set.

The paper is organized as follows. In the first section, we propose an analysis of the recent literature on multicriteria sorting methods. Section 2 presents the NCS model. Inv-NCS, the problem of learning the parameters of NCS from assignment examples is defined in Section 3. In Section 4, we present the two SAT formulations for Inv-NCS. In Section 5, we extend SAT formulations with MaxSAT language, and Section 6 describes the empirical test design, the experimental results and a discussion. A final section groups conclusions and avenues for further research.

The main contributions of the paper concern the extension of the separation-based SAT formulation to the case of more than two categories (the end of Section 4), the extension of the SAT formulations to MaxSAT to account for noisy input (Section 5), and empirical results providing insights on how our methods behave on actual data sets (Section 6). However, we also provide a comprehensive description of the NCS models (Section 2), so as to provide a self-contained text, as well as a brief survey of the recent literature concerning the elicitation of MCDA sorting models (Section 1).

1. Recent literature on multiple criteria sorting methods

Many multicriteria sorting models have been proposed in the literature (see Doumpos & Zopounidis, 2002 for an overview). These multicriteria sorting models can be distinguished according to the way they model preferences: (*i*) the ones that model preferences using a multi-attribute value function (e.g., Corrente, Doumpos, Greco, Słowiński, & Zopounidis, 2017; Devaud, Groussaud, & Jacquet-Lagreze, 1980; Köksalan & Ozpeynirci, 2009; Marichal, Meyer, & Roubens, 2005; Siskos, Grigoroudis, & Matsatsinis, 2016), (*ii*) the ones that model preferences using outranking relations (e.g., Almeida-Dias, Figueira, & Roy, 2012; Fernández, Figueira, Navarro, & Roy, 2017; Kadziński, Tervonen, & Figueira, 2015; Perny, 1998; Roy, 1991), and (*iii*) those which represent preferences using if-then rules (e.g., Błaszczyński, Greco, & Słowiński,

2007; Greco, Matarazzo, & Slowinski, 2001; Kadziński, Greco, & Słowiński, 2014; Rudin & Ertekin, 2018). Recently Kadzinski, Ghaderi, & Dabrowski (2020) proposed a method which corresponds to a hybridization of value-based and rule-based approaches.

Learning preference models from preference data to faithfully represent the DM's judgment has been considered since several decades in the literature. In the context of MCDA, a well-known example of such an approach is the UTA method proposed in Jacquet-Lagreze & Siskos (1982) in the case of an additive multicriteria value model. Learning an Electre Tri model (the initial multicriteria sorting procedure from which NCS was formalized, see Roy (1991)) from assignment examples was initially formulated using non-linear programming in Mousseau & Słowiński (1998).

A significant number of authors did focus on the robustness of sorting results. Some approaches are based on robust ordinal regression, e.g. Greco, Mousseau, & Slowinski (2010), while some others focus on a stochastic approach, e.g. Tervonen, Figueira, Lahdelma, Almeida Dias, & Salminen (2009). Another concern which emerged from the literature on sorting methods concerns the ability to explain the result to the decision maker in order to reinforce her trust (see Belahcene, Labreuche, Maudet, Mousseau, & Ouerdane, 2017a; Belahcene et al., 2018b; Labreuche, 2011).

Since one decade, the literature on outranking-based sorting has widely expanded: based on the Electre Tri method (Roy, 1991), new sorting methods have been proposed defining categories using one or several limit profiles (see Fernández, Figueira, & Navarro, 2019; Fernández et al., 2017), one or several central profiles (see Almeida-Dias, Figueira, & Roy, 2010; Almeida-Dias et al., 2012; Kadziński et al., 2015). In parallel, several theoretical and axiomatic works have contributed to a better understanding of these methods (see Bouyssou & Marchant, 2007a; Bouyssou & Marchant, 2007b; Bouyssou, Marchant, & Pirlot, 2020).

An additional significant advance in the literature concerns the new proposals in incremental elicitation of sorting methods (see Benabbou, Perny, & Viappiani, 2016; Benabbou, Perny, & Viappiani, 2017; Kadzinski & Ciomek, 2021). These works propose a strategy to iteratively select questions to be asked to the decision maker in order to limit the number of questions.

Several new sorting methods allow to cope with possible interaction between criteria (see e.g. Fallah Tehrani, Cheng, & Hüllermeier, 2011; Liu, Kadzinski, Liao, & Mao, 2021). The possibility to represent preferences in a hierarchical structure of criteria has also been considered (see Arcidiacono, Corrente, & Greco, 2021). Some authors also consider non monotone preferences in sorting methods (see Liu, Liao, Kadziński, & Słowiński, 2019; Liu, Liao, Mao, Wang, & Kadzinski, 2020; Minoungou, Mousseau, Ouerdane, & Scotton, 2020). These features enable sorting methods to account for more flexible preferences.

Another important trend in the sorting literature concerns the cross-fertilization between the field of MCDA sorting and preference learning (Furnkranz & Hullermeier, 2011). These two communities have now a common conference DA2PL (from Decision Analysis to Preference Learning) which takes place every second year since 2012. Among preference learning related work, one can cite (Fallah Tehrani et al., 2011; Liu et al., 2021; Liu et al., 2019; Liu et al., 2020; Rudin & Ertekin, 2018; Sobrie, 2016). It should be noted that in this perspective, medical applications have been a fruitful application domain (see e.g Sobrie, Lazouni, Mahmoudi, Mousseau, & Pirlot, 2016; Sokolovska, Chevaleyre, & Zucker, 2018; Ustun & Rudin, 2016).

Coming back specifically to outranking-based sorting, previous works have proposed approaches to learn the parameters of an MR-Sort model (specific case of an NCS model in which the set of sufficient coalitions of criteria are defined using additive weights) based on a learning set. This MIP formulation minimizes the 0/1

loss, i.e. searches for a model that is compatible with as many examples as possible. Such MIP based exact approach has been extended to NCS 2-additive models (Sobrie et al., 2015). However, experimental results showed that such MIP approach becomes computationally prohibitive with a large number of assignments (learning an MR-Sort model with 100 alternatives, 5 criteria and 3 categories involve a MIP with 1100 binary variables, and the computing time exceeds 100 s).

To cope with the computational burden, a heuristic approach has been proposed to learn an MR-Sort model from assignment examples by Sobrie (2016) and Sobrie, Mousseau, & Pirlot (2019) which can handle large datasets, but losing optimality guaranty. More recently Belahcene et al. (2018a) defined a Boolean satisfiability formulation of Inv-NCS, which keeps optimality guarantee while enabling computations even for real-size datasets. In this paper, we continue and extend this work.

2. Non-compensatory sorting models

This section is devoted to the presentation of the Non-Compensatory Sorting model, introduced in Bouyssou & Marchant (2007a,b).

2.1. Basic notations

Multicriteria sorting aims at assigning alternatives to one of the predefined ordered categories $C^1 \prec \ldots \prec C^p$. All alternatives in a set A are evaluated on n criteria, $\mathcal{N} = \{1, 2, \ldots, n\}$; hence, an alternative $a \in A$ is characterized by its evaluation vector $(a_1, \ldots, a_n)$, with $a_i \in \mathbb{X}_i$ denoting its evaluation on criterion i. Each criterion is equipped with a weak preference relation $\succsim_i$ defined on $\mathbb{X}_i$. We assume, without loss of generality, that the preference on each criterion increases with the evaluation (the greater, the better). We denote by $\mathbb{X} = \prod_{i \in \mathcal{N}} \mathbb{X}_i$ the cartesian product of evaluation scales.

We recall the definitions of an upset and the upper closure of a subset w.r.t. a binary relation:

Definition 2.1. (Upset and upper closure). Let $\mathcal{A}$ be a set and $\mathcal{R}$ a binary relation on $\mathcal{A}$. An upset of $(\mathcal{A}, \mathcal{R})$ is a subset $\mathcal{B} \subseteq \mathcal{A}$ such that $\forall a \in \mathcal{A}, \forall b \in \mathcal{B}, a\mathcal{R}b \Rightarrow a \in \mathcal{B}$. The upper closure of a subset of $(\mathcal{A}, \mathcal{R})$ is the smallest upset of $(\mathcal{A}, \mathcal{R})$ containing it: $\forall \mathcal{B} \subseteq \mathcal{A}, cl^{\mathcal{R}}_{\mathcal{A}}(\mathcal{B}) := \{a \in \mathcal{A} : \exists b \in \mathcal{B}\ a\mathcal{R}b\}$

2.2. Sorting into two categories

In the Non-Compensatory Sorting model (NCS), the boundaries between categories are defined by limiting profiles. Therefore, a single profile corresponds to the case where alternatives are sorted between two ordered categories that we label as Good and Bad. A pair of parameters describe a specific sorting procedure:

- a limiting profile $b \equiv \langle b_i \rangle_{i \in \mathcal{N}}$ that defines, according to each criterion $i \in \mathcal{N}$, an upper set $\mathcal{A}_i \subset \mathbb{X}_i$ of approved values at least as good as b_i (and, by contrast, a lower set $\mathbb{X} \setminus \mathcal{A}_i \subset \mathbb{X}_i$ of disapproved values strictly worse than b_i), and
- a set $\mathcal{T}$ of sufficient coalitions of criteria, which satisfies monotonicity with respect to inclusion.

These notions are combined into the following assignment rule:

$$\forall x \in \mathbb{X}, \quad x \in \text{Good} \iff \{i \in \mathcal{N} : x_i \succsim_i b_i\} \in \mathcal{T}$$

An alternative is considered as Good if, and only if, it is better than the limiting profile b according to a sufficient coalition of criteria.

Table 1
Performance table.

model	cost	acceleration	braking	road holding
m_1	16 973€	29.0 sec.	2.66	2.5
m_2	18 342€	30.7 sec.	2.33	3
m_3	15 335€	30.2 sec.	2	2.5
m_4	18 971€	28.0 sec.	2.33	2
m_5	17 537€	28.3 sec.	2.33	2.75
m_6	15 131€	29.7 sec.	1.66	1.75

Table 2
Limiting profiles.

profile	cost	acceleration	braking	road holding
$b^{1\star}$	17 250€	30.0 sec.	2.2	1.9
$b^{2\star}$	15 500€	28.8 sec.	2.5	2.6

2.3. Sorting into multiple categories

With p categories, the parameter space is extended accordingly, with approved sets $\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N},\ k \in [2.p]}$ defined by a set of limiting profiles $\langle b_i^k \rangle_{i \in \mathcal{N},\ k \in [2.p]}$ and sufficient coalitions $\langle \mathcal{T}^k \rangle_{k \in [2.p]}$ declined per boundary.

The ordering of the categories $\{C^1 \prec \ldots \prec C^p\}$ translates into a nesting of the sufficient coalitions: $\forall k \in [2.p]$, $\mathcal{T}^k$ is an upset of $(2^{\mathcal{N}}, \subseteq)$ and $\mathcal{T}^2 \supseteq \cdots \supseteq \mathcal{T}^p$, and also a nesting of the approved sets: $\forall i \in \mathcal{N}, \forall k \in [2.p]$, $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \succsim_i)$ and $\mathcal{A}_i^2 \supseteq \cdots \supseteq \mathcal{A}_i^p$.

These tuples of parameters are augmented on both ends with trivial values: $\mathcal{T}^1 = \mathcal{P}(\mathcal{N})$, $\mathcal{T}^{p+1} = \emptyset$, and $\forall i \in \mathcal{N}$, $\mathcal{A}_i^1 = \mathbb{X}$, $\mathcal{A}_i^{p+1} = \emptyset$. With $\omega = (\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N},\ k \in [2.p]}, \langle \mathcal{T}^k \rangle_{k \in [2.p]})$, Bouyssou & Marchant (2007b) defines the sorting function NCS_ω from $\mathbb{X}$ to $\{C^1 \prec \ldots \prec C^p\}$ with the *Non-Compensatory Sorting rule*:

$$NCS_\omega(x) = C^k \Leftrightarrow \begin{cases} \{i \in \mathcal{N} : x \in \mathcal{A}_i^k\} & \in \mathcal{T}^k \\ \text{and } \{i \in \mathcal{N} : x \in \mathcal{A}_i^{k+1}\} & \notin \mathcal{T}^{k+1} \end{cases} \quad (1)$$

Note that Bouyssou & Marchant (2007a,b) define a broader class of sorting method which includes vetoes which makes it possible for a single criterion to forbid the assignment to a class. Throughout this paper, we only consider NCS without veto; therefore, we should formally write NCS without veto all along with the paper. However, to facilitate the reading, we choose to write NCS even if we consider NCS models without a veto.

2.4. An illustrative example

A journalist prepares a car review for a forthcoming issue. She considers a number of popular car models and wants to sort them to present a sample of cars "selected for you by the editorial board" to the readers. This selection is based on four criteria: cost (€), acceleration (time, in seconds, to reach 100 km/h from full stop – lower is better), braking power and road holding, both measured on a qualitative scale ranging from 1 (lowest performance) to 4 (best performance). The performances of the six models are described in Table 1.

In order to assign these models to a category among $C^{1\star}$ (average) $\prec C^{2\star}$ (good) $\prec C^{3\star}$ (excellent), the journalist considers an NCS model:

- The attributes of each model are sorted between average ($\star$/ 🟥), good ($\star\star$/ 🟨) and excellent ($\star\star\star$/ 🟩) by comparison to the profiles given in Table 2. The resulting labeling of the six alternatives according to each criterion is depicted in Fig. 1 and Table 3.

- These appreciations are then aggregated by the following rule: *an alternative is categorized good or excellent if it is good or*
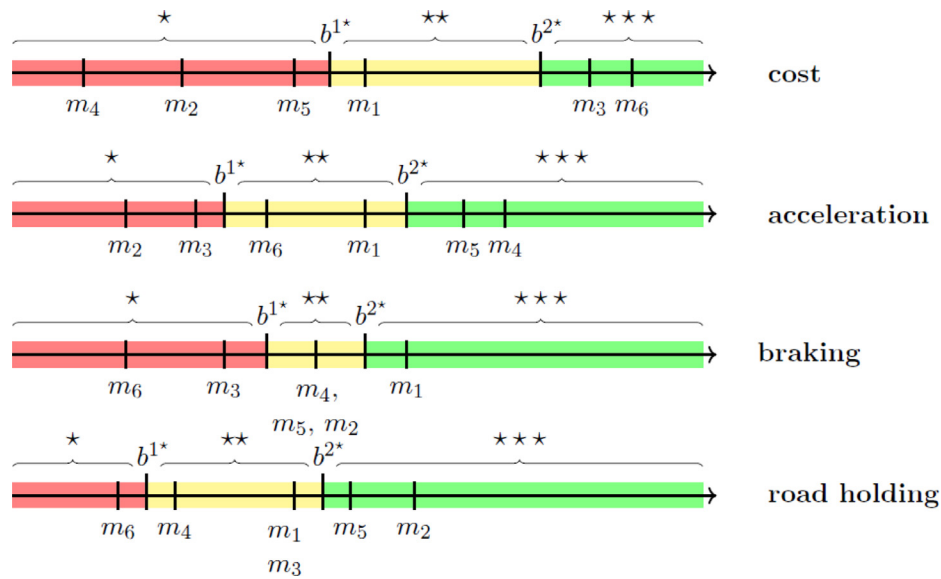
Fig. 1. Representation of performances w.r.t. category limits.

Table 3
Categorization of performances.

model	cost	acceleration	braking	road holding
m_1	★★	★★	★★★	★★
m_2	★	★	★★	★★★
m_3	★★★	★	★	★★
m_4	★	★★★	★★	★★
m_5	★	★★★	★★	★★★
m_6	★★★	★★	★	★

Table 4
Alternative assignments.

Alternatives	m_1	m_2	m_3	m_4	m_5	m_6
Assignment	★★	★	★★	★★	★★★	★

Fig. 2. Sufficient (green/thick-bordered) and insufficient (red/thin-bordered) coalitions of criteria. Arrows denote coalition strength. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

excellent on cost or acceleration, and good or excellent on braking or road holding. It is categorized excellent if it is excellent on cost or acceleration, and excellent on braking or road holding. Being excellent on some criterion does not really help to be considered good overall, as expected from a Non-Compensatory model. Sufficient coalitions are represented on Fig. 2. Finally, the model yields the assignments presented in Table 4.

2.5. Variants of the NCS model

In this section, we mention a number of variants of the Non-Compensatory Sorting model that can be found in the literature. Note that Bouyssou & Marchant (2007a,b) define the NCS class of sorting method, which includes the possibility of vetoes. In this paper, we only consider NCS without veto, but it should be highlighted that the broader class of NCS model can include vetoes, as depicted in Fig. 3. Among NCS models without veto, there exist variants corresponding to simplifications of the model, with additional assumptions that restrict the parameters—limiting profiles and sufficient coalitions—either explicitly or implicitly.

The set of preference parameters – all the pairs $(\langle \mathcal{A} \rangle, \langle \mathcal{T} \rangle)$ can be considered too wide and too unwieldy for practical use in the context of a decision aiding process. Therefore, following Bouyssou & Marchant (2007b), one may consider to explicitly restrict either the sequence of limiting profiles, or the sequence of sufficient coalitions:

- U^C-NCS: *Non-Compensatory Sorting with a unique set of sufficient coalitions*: $\mathcal{T}^2 = \cdots = \mathcal{T}^p$;
- U^B-NCS: *Non-Compensatory Sorting with a unique limiting profile* $b^2 = \cdots = b^p$ or, equivalently, $\forall i \in \mathcal{N}, \mathcal{A}_i^2 = \cdots = \mathcal{A}_i^p$.

It worth to be noted that an NCS model which is in U^C-NCS and U^B-NCS simultaneously corresponds necessarily to a model with two categories (cf. the intersection colored in blue in Fig. 3).

Another simplifying assumption consists in representing sufficient coalitions additively in an analogy to a voting setting: each criterion $i \in \mathcal{N}$ is assigned with a *voting power* $w_i \geq 0$ so that a given coalition of criteria $B \subseteq \mathcal{N}$ is deemed sufficient if, and only if, its combined voting power $\sum_{i \in B} w_i$ is greater than a given
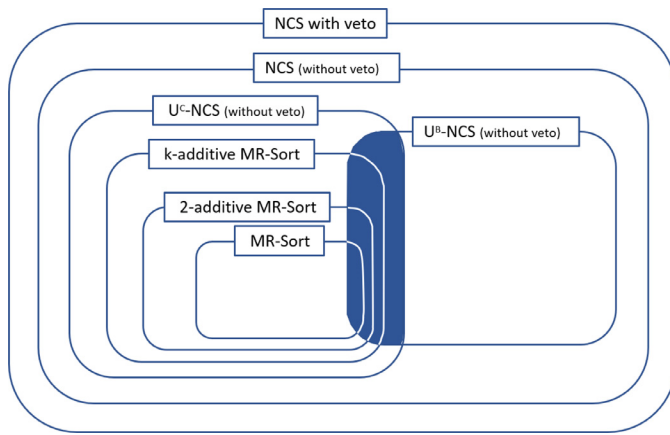
Fig. 3. Variants of the NCS model.

qualification threshold λ.

$$\exists \lambda, \langle w_i \rangle_{i \in \mathcal{N}} \in [0,1] : \ \forall B \subseteq \mathcal{N}, B \in \mathcal{T} \iff \sum_{i \in B} w_i \geq \lambda. \tag{2}$$

With this rule, the sufficient coalitions are represented in a compact form which is more amenable to linear programming. This additive version of U^C-NCS is frequently called MR-Sort (for *majority rule sorting*) in the literature (see, e.g. Leroy et al., 2011).

A more general way to describe possible interactions between criteria coalitions is to represent these coalitions using a capacity $\mu : 2^{\mathcal{N}} \mapsto [0,1]$, with $\mu(\emptyset) = 0$, $\mu(\mathcal{N}) = 1$, and $\mu(B) \geq \mu(A)$, for all $A \subseteq B \subseteq \mathcal{N}$. The Möbius transform allows to express a capacity μ in another form: $\mu(A) = \sum_{B \subseteq A} m(B), \forall A \subseteq \mathcal{N}$ with $m(B) = \sum_{C \subseteq B} (-1)^{|B| - |C|} \mu(C)$. The value $m(B)$ can be interpreted as the weight that is allocated to B as a whole. A capacity can be defined directly by its Möbius transform also called Möbius interaction. A Möbius interaction or Möbius mass m is a set function $m : 2^{\mathcal{N}} \mapsto [-1,1]$ satisfying the hereafter conditions which guarantee that μ is monotone (see Chateauneuf & Jaffray, 1987):

$$\sum_{j \in K \subseteq J \cup \{j\}} m(K) \geq 0, \forall j \in \mathcal{N}, \forall J \subseteq \mathcal{N} \setminus \{j\} \text{ and } \sum_{K \subseteq \mathcal{N}} m(K) = 1.$$

Using such representation, it is possible to consider 2-additive (k-additive, resp.) capacities for which all the interactions involving more than 2 (k, resp.) criteria are equal to zero. 2-additive and k-additive MR-Sort (2-additive and k-additive U^C-NCS) are represented in Fig. 3 (although not depicted, it is also possible to consider k-additive U^B-NCS).

3. Learning an NCS model from data

For a given decision situation, assuming the NCS model is relevant to structure the decision maker's preferences, what should be the parameters' values to fully specify the NCS model that corresponds to the decision-maker (DM) viewpoint? An option would be to simply ask the decision-maker to describe, to her best knowledge, the limit profiles between categories and to enumerate the minimal sufficient coalitions. To get this information as quickly and reliably as possible, an analyst could make good use of the *model-based elicitation strategy* described in Belahcene, Mousseau, Pirlot, & Sobrie (2017b), as it permits to obtain these parameters by asking the decision-maker to only provide holistic preference judgment – should some (fictitious) alternative be assigned to some category – and build the shortest questionnaire.

We opt for a more indirect setup, close to a machine learning paradigm (Furnkranz & Hullermeier, 2011), where a set of reference assignments is given and assumed to describe the decision-maker's

point of view, and the aim is to *extend* these assignments with an NCS model. In this context, we usually refer to an *assignment* as a function mapping a subset of *reference alternatives* $\mathbb{X}^{\star} \subset \mathbb{X}$ to the ordered set of categories $C^1 \prec \cdots \prec C^p$. These reference alternatives highlight values of interest on each criterion $i \in \mathcal{N}$, $\mathbb{X}_i^{\star} = \bigcup_{x \in \mathbb{X}^{\star}} \{x_i\}$. We refer to the problem of finding suitable preference parameters specifying a Non-Compensatory Sorting model by Inv-NCS.

3.1. NCS and Inv-NCS

Instances An *instance* of the Inv-NCS problem is a sextuple $(\mathcal{N}, \mathbb{X}, \langle \succsim_i \rangle_{i \in \mathcal{N}}, \mathbb{X}^{\star}, \{C^1 \prec \ldots \prec C^p\}, \alpha)$ where:

- $\mathcal{N}$ is a set of criteria;
- $\mathbb{X}$ is a set of *alternatives*;
- $\langle \succsim_i \rangle_{i \in \mathcal{N}} \in \mathbb{X}^2$ are *preferences* on criterion i, $i \in \mathcal{N}$, $\succsim_i \subset \mathbb{X}^2$ is a total pre-ordering of alternatives according to this criterion;
- $\mathbb{X}^{\star} \subset \mathbb{X}$ is a finite set of *reference alternatives*;
- $\{C^1 \prec \ldots \prec C^p\}$ is a finite set of *categories* totally ordered by *exigence*. We denote $C^{\geq k}$ (resp. $C^{>k}, C^{\leq k}, C^{<k}$) the category interval $\{C^k \prec \cdots \prec C^p\}$ (resp. $\{C^{k+1} \prec \cdots \prec C^p\}$, $\{C^1 \prec \ldots C^k\}$, $\{C^1 \prec \cdots \prec C^{k-1}\}$);
- $\alpha : \mathbb{X}^{\star} \to \{C^1 \prec \ldots \prec C^p\}$ is an *assignment* of the reference alternatives to the categories. Therefore, 'α^{-1}' is the associated inverse function i.e. for a given category C^h, $\alpha^{-1}(C^h) = \{x \in \mathbb{X}^{\star} : x \in C^h\}$. For any comparison operator $\Delta \in \{\succ, \succeq, \prec, \preceq\}$, we also denote $\alpha^{-1}(C^{\Delta h}) := \{x \in \mathbb{X}^{\star} : x \in C^k, \ C^k \Delta \ C^h\}$.

When referring to an instance, we often shorten this sextuple as 'α'.

Parameters Given a context, a *parameter* ω of the NCS model is a couple $(\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, \ k \in [2.p]}, \langle \mathcal{T}^k \rangle_{k \in [2.p]})$, where the *sufficient coalitions* satisfy: $\forall k \in [2.p]$, $\mathcal{T}^k$ is an upset of $(2^{\mathcal{N}}, \subseteq)$, and $\mathcal{T}^2 \supseteq \cdots \supseteq \mathcal{T}^p$; and the *approved sets* satisfy $\forall i \in \mathcal{N}$, $\forall k \in [2.p]$, $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \preceq_i)$ and $\mathcal{A}_i^2 \supseteq \cdots \supseteq \mathcal{A}_i^p$.

Sorting rule Given a parameter $\omega = (\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, \ k \in [2.p]}, \langle \mathcal{T}^k \rangle_{k \in [2.p]})$, augmented with trivial values $\mathcal{T}^1 := \mathcal{P}(\mathcal{N})$, $T^{p+1} := \emptyset$, $\forall i \in \mathcal{N}$, $\mathcal{A}_i^2 = \mathbb{X}$, $\mathcal{A}_i^{p+1} = \emptyset$, NCS_{ω} is the function from $\mathbb{X}$ to $\{C^1 \prec \ldots \prec C^p\}$ satisfying:

$$NCS_{\omega}(x) = C^k \Leftrightarrow \begin{cases} \forall k' \leq k, & \{i \in \mathcal{N} : x \in \mathcal{A}_i^{k'}\} \in \mathcal{T}^{k'} \text{ and} \\ \forall k' > k, & \{i \in \mathcal{N} : x \in \mathcal{A}_i^{k'}\} \notin \mathcal{T}^{k'}. \end{cases} \tag{3}$$

This rule can be equivalently written as follows:

$$NCS_{\omega}(x) \in C^{\geq k} \Leftrightarrow \{i \in \mathcal{N} : x \in \mathcal{A}_i^k\} \in \mathcal{T}^k. \tag{4}$$

Solutions Given a context, a *solution* of the instance α of the Inv-NCS problem is a parameter ω of the NCS model such that $\forall x \in \mathbb{X}^{\star}$, $\alpha(x) = NCS_{\omega}(x)$.

3.2. Literature related to Inv-NCS

Learning preference models from preference data to faithfully represent the DM judgment has been considered since several decades in the literature. In the context of MCDA, a well-known example of such an approach is the UTA method proposed in Jacquet-Lagreze & Siskos (1982) in the case of an additive multicriteria value model. Learning an Electre Tri model (the initial multicriteria sorting procedure from which NCS was formalized, see Roy (1991)) from assignment examples was initially formulated using non-linear programming in Mousseau & Słowiński (1998). A mixed-integer linear formulation was proposed by Leroy et al. (2011) to learn an additive majority rule sorting model (MR-Sort: additive NCS without veto) from a dataset; however, these approaches were not able to handle datasets corresponding to real-world problems. Recently, Kadzinski & Martyn (2020) proposed an

enriched framework to elicit and Electre Tri B model and analyze its results.

To cope with the computational burden, a heuristic approach has been proposed to learn an MR-Sort model from assignment examples by (Sobrie, 2016; Sobrie et al., 2019) which can handle large datasets, but losing optimality guaranty. More recently Belahcene et al. (2018a) defined a Boolean satisfiability formulation of Inv-NCS, which keeps optimality guarantee while enabling computations even for real-size datasets. In this paper, we continue and extend this work.

4. Boolean satisfiability formulations for the Inv-NCS problem

This section is devoted to the presentation of two formulations of the inverse Non-Compensatory Sorting problem, first described respectively in Belahcene et al. (2018a) and Belahcene et al. (2018b), in the framework of Boolean satisfiability. They reduce the problem of finding the parameters of an NCS model faithfully reproducing a given assignment of alternatives to categories to the SAT problem of finding an assignment of Boolean variables that verify a given propositional formula written in conjunctive normal form.[1]

The two formulas stem from different representation strategies. One, detailed in Section 4.1 and introduced in Belahcene et al. (2018a), establishes a bijection between the parameter space of the NCS model and the valuations of the propositional variables, and therefore introduces a number of variables that is exponential in the number of criteria. The other is detailed in Section 4.3 and was introduced in Belahcene et al. (2018b). It leverages a powerful representation theorem, detailed in Section 4.2, that allows keeping implicit the set of coalitions, by introducing the notion of *pairwise separation*, formulated using pairs of alternatives given in the assignment.

Appendix A complements this section by providing previously unpublished formulations for the case where there are more than two categories, including the variants with a unique profile or a unique set of sufficient coalitions described in Section 2.5.

4.1. A SAT formulation for Inv-NCS based on coalitions

This section describes and extends a SAT formulation for Inv-NCS initially given in Belahcene et al. (2018a). We provide here an informal presentation of the approach; formal justification can be found in Belahcene et al. (2018a). The formulation Φ_α^C yielded by the encoding presented in this section is based on an explicit representation of the parameter space of the Non-Compensatory Sorting model—each the pairs are composed of a sequence of approved sets and a sequence of sufficient coalitions.

The explicit representation requires involving two families of binary variables. The first family defines the approved sets according to the set of criteria such that for a given alternative, level and criterion, the associated variable equals 1 if and only if the alternative is approved at the considered level according to the considered criterion. The second family of binary variables uniquely specifies the set of sufficient coalitions for each level i.e. given a coalition of criteria, the associated variable equals 1 if and only if the coalition is sufficient. The SAT formulation based on coalitions aims at learning both NCS parameters ($\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N},\ k \in [2.p]}$, $\langle \mathcal{T}^k \rangle_{k \in [2.p]}$) from a set of assignment examples, thus, two types of clauses are considered. The first type of clauses defines these parameters and reproduces

the structural conditions i.e.: the monotony of scales, approved sets and sufficient coalitions sets are ordered by inclusion and for each level the corresponding set sufficient coalitions is monotone by inclusion. The second type of clauses ensures the restoration of the assignment examples.

Variables The Boolean function Φ_α^C operates on two types of variables:

- 'a' variables, indexed by a criterion $i \in \mathcal{N}$, an exigence level $k \in [2.p]$ and a reference value $x \in \mathbb{X}^\star$, represent the approved sets $\mathcal{A}_i^k$, with the following semantic: $a_{i,k,x} = 1 \Leftrightarrow x \in \mathcal{A}_i^k$ i.e. x is approved at level k according to i;
- 't' variables, indexed by a coalition of criteria $B \subseteq \mathcal{N}$ and an exigence level $k \in [2.p]$, represent the sufficient coalitions $\mathcal{T}^k$, with the following semantic: $t_{B,k} = 1 \Leftrightarrow B \in \mathcal{T}^k$ i.e. the coalition B is sufficient at level k;

Clauses For a boolean function written in conjunctive normal form, the clauses are *constraints* that must be satisfied simultaneously by any antecedent of 1. The formulation Φ_α^C is built using six types of clauses:

- Clauses ϕ_α^{C1} ensure that each approved set $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}^\star, \precsim_i)$: if for a criterion i and an exigence value k, the value x is approved, then any value $x' \succsim_i x$ must also be approved.
- Clauses ϕ_α^{C2} ensure that approved sets are ordered by a set inclusion according to their exigence level: if an alternative x is approved at exigence level k according to the criterion i, it should also be approved at exigence level $k' < k$.
- Clauses ϕ_α^{C3} ensure that each set of sufficient coalitions $\mathcal{T}$ is an upset for inclusion: if a coalition B is deemed sufficient at exigence level k, then a stronger coalition $B' \supset B$ should also be deemed sufficient at this level.
- Clauses ϕ_α^{C4} ensure that a set of sufficient coalitions are ordered by inclusion according to their exigence level: if a coalition B is deemed insufficient at exigence level k, it should also be at any level $k' > k$.
- Clauses ϕ_α^{C5} ensure that each alternative is not approved by a sufficient coalition of criteria at an exigence level above the one corresponding to its assigned category.
- Clauses ϕ_α^{C6} ensure that each alternative is approved by a sufficient coalition of criteria at an exigence level corresponding to its assignment.

Definition 4.1. Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$, the boolean function Φ_α^C with variables $\langle a_{i,k,x} \rangle_{i \in \mathcal{N},\ k \in [2.p],\ x \in \mathbb{X}^\star}$ and $\langle t_{B,k} \rangle_{B \subseteq \mathcal{N},\ k \in [2.p]}$, is defined as the conjunction of clauses:

$$\Phi_\alpha^C = \phi_\alpha^{C1} \wedge \phi_\alpha^{C2} \wedge \phi_\alpha^{C3} \wedge \phi_\alpha^{C4} \wedge \phi_\alpha^{C5} \wedge \phi_\alpha^{C6}$$

$$\phi_\alpha^{C1} = \bigwedge_{i \in \mathcal{N},\ k \in [2.p]} \bigwedge_{x' \succsim_i x \in \mathbb{X}^\star} (a_{i,k,x'} \vee \neg a_{i,k,x})$$

$$\phi_\alpha^{C2} = \bigwedge_{i \in \mathcal{N},\ k < k' \in [2.p],\ x \in \mathbb{X}^\star} (a_{i,k,x} \vee \neg a_{i,k',x})$$

$$\phi_\alpha^{C3} = \bigwedge_{B \subset B' \subseteq \mathcal{N},\ k \in [2.p]} (t_{B',k} \vee \neg t_{B,k})$$

$$\phi_\alpha^{C4} = \bigwedge_{B \subseteq \mathcal{N},\ k < k' \in [2.p]} (t_{B,k} \vee \neg t_{B,k'})$$

$$\phi_\alpha^{C5} = \bigwedge_{B \subseteq \mathcal{N},\ k \in [2.p]} \bigwedge_{x \in \alpha^{-1}(C^{k-1})} \left(\bigvee_{i \in B} \neg a_{i,k,x} \vee \neg t_{B,k} \right)$$

$$\phi_\alpha^{C6} = \bigwedge_{B \subseteq \mathcal{N},\ k \in [2.p]} \bigwedge_{x \in \alpha^{-1}(C^k)} \left(\bigvee_{i \in B} a_{i,k,x} \vee t_{\mathcal{N} \setminus B,k} \right)$$

Written as such, clauses of Φ_α^C are highly redundant, possibly threatening computational efficiency.[2] Instead, it is sufficient

---

[1] For the convenience of EJOR readers, who might be more accustomed to the formalism of Mathematical Programming, we treat SAT as the tiny subset of MP where the variables are restricted to the $\{0,1\}$ domain, the objective function is null, and the constraints are limited to linear forms of the type $\sum_{i \in C_j^+} x_i + \sum_{i \in C_j^-} (1 - x_i) \geq 1$, corresponding to the clause $\bigvee_{i \in C_j^+} x_i \vee \bigvee_{i \in C_j^-} \neg x_i$.

[2] Even though SAT solvers often perform better on reasonably overconstrained problems.

to consider clauses where ordered elements in a pair are adjacent to each other.

Model variants As discussed in Section 2.5, the NCS model has many variants. Φ^C_α can easily be modified to account for two popular restrictions of the model:

- U^B-NCS (Unique profiles): Drop the index k concerning the exigence level for the 'a' variables, ignore the conjunction over exigence levels for clauses ϕ^{C1}_α, and ignore clauses ϕ^{C2}_α altogether;
- U^C-NCS (Unique set of sufficient coalitions): Drop the index k concerning the exigence level for the 't' variables, ignore the conjunction over exigence levels for clauses ϕ^{C3}_α and ignore clauses, ϕ^{C4}_α altogether.

4.2. A characterization based on pairwise separation

4.2.1. The case of two categories

The problem of finding simultaneously the sets of accepted values of the criteria and the sets of sufficient coalitions has been considered computationally difficult from the onset. In this light, the assumption of an additive representation of sufficient coalitions with the *majority rule* can be considered as a convenient way to keep the search somewhat tractable[3] Indeed, when the accepted values are known, finding the parameters (the voting power of each criterion and the qualification threshold) of a suitable majority rule becomes a mere linear program with continuous variables and can be solved in polynomial time. It is possible to represent the NCS model with two categories in the MAVT paradigm, using full-fledged $|\mathcal{N}|$-ary capacities, but the corresponding linear program requires $2^{|\mathcal{N}|}$ variables. This approach is deceptively difficult, though, and we shall see that, from the viewpoint of Computer Theory, Inv-NCS is actually not more difficult than its restriction to additive coalitions. This result comes from a simple series of observations. In the following, we suppose given a set of reference alternatives $\mathbb{X}^\star$, an assignment $\alpha : \mathbb{X}^\star \to \{$ Good, Bad$\}$, and a tuple of accepted values $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{|\mathcal{N}|}$ such that, for each point of view $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$. *Observably sufficient and insufficient coalitions* Consider the sets of coalitions defined by

$$\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^\supseteq_{\mathcal{P}(\mathcal{N})}\left(\bigcup_{g \in \alpha^{-1}(Good)} \{i \in \mathcal{N} : g \in \mathcal{A}_i\} \right), \tag{5}$$

$$\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^\subseteq_{\mathcal{P}(\mathcal{N})}\left(\bigcup_{b \in \alpha^{-1}(Bad)} \{i \in \mathcal{N} : b \in \mathcal{A}_i\} \right). \tag{6}$$

Any coalition in $\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is a superset of the set of criteria according to which some Good alternative is accepted, and should, therefore, be accepted. Thus, $\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is a *lower bound* of the set of sufficient coalitions for any solution of Inv-NCS. Conversely, any coalition in $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is a subset of the set of criteria according to which some Badalternative is accepted, and should, therefore, be rejected. Thus, $\mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is an *upper bound* of the set of sufficient coalitions for any solution of Inv-NCS. *Characterization of solutions of Inv-NCS* The parameter $(\langle \mathcal{A}_i \rangle, \mathcal{T})$ is a solution of the instance α of Inv-NCS if and only if:

$$\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha) \subseteq \mathcal{T} \subseteq \mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) \tag{7}$$

Remarkably, this equation allows to characterize the positive instances of Inv-NCS without referring to the set of sufficient of coalitions of a solution, solely by checking if the sets $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ are disjoint. This leads to the following elegant and efficient characterization, based on the notion of *pairwise separation*.

Theorem 4.1. *An assignment α of alternatives to categories can be represented in the Non-Compensatory Sorting model if, and only if, there is a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{|\mathcal{N}|}$ such that:*

1. *(Upset): for each point of view $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$; and*
2. *(Pairwise separation): for each pair of alternatives $(g, b) \in \alpha^{-1}($ Good$) \times \alpha^{-1}($ Bad$)$, there is at least one point of view $i \in \mathcal{N}$ such that $g \in \mathcal{A}_i$ and $b \notin \mathcal{A}_i$.*

This theorem provides a polynomial certificate for the positive instances of the Inv-NCS problem, thus proving its membership of the *NP* complexity class as a corollary. Proofs of Theorem 4.1, and of the NP-hardness of Inv-NCS can be found in Belahcene et al. (2018b). The extension of this characterization to any number of categories is straightforward and is presented in the following section and Appendix A.

4.2.2. The case of more than two categories

The case where there are $p > 2$ categories $\{C^1 \prec \ldots \prec C^p\}$ requires a few adaptations of the formulation. It relies mostly on the fact that an NCS model with p categories is, informally, the combination of $p - 1$ NCS models with two categories whose parameters satisfy the nesting conditions on the sufficient coalitions of criteria and the accepted values.

Given an assignment α and an exigence level $k \in [2.p]$, we define the set of alternatives assigned to categories better than and including C^k denoted $C^{\succeq k}$ and the set of alternatives assigned to categories worse than C^k denoted $C^{\prec k}$ as:

$$C^{\succeq k} = \bigcup_{h \in [k.p]} C^h; \quad C^{\prec k} = \bigcup_{h \in [2.k-1]} C^h$$

We extend Eqs. (5) and (6) so that, at a given exigence level k, observably sufficient coalitions account for "good" alternatives in $C^{\succeq k}$ and observably insufficient coalitions account for "bad" alternatives in $C^{\prec k}$.

Definition 4.2. (Observed sufficient and insufficient coalitions given approved sets). Given an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$, approved sets $\langle \mathcal{A}^k_i \rangle_{i \in \mathcal{N}, \ k \in [2.p]}$ such that $\mathcal{A}^k_i$ is an upset of $(\mathbb{X}_i, \precsim_i)$ and $\mathcal{A}^2_i \supseteq \cdots \supseteq \mathcal{A}^p_i$, we note, for any exigence level $k \in [2.p]$:

$$\mathcal{S}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha) = Cl^\supseteq_{\mathcal{P}(\mathcal{N})}(\bigcup_{g \in \alpha^{-1}(C^{\succeq k})} \{i \in \mathcal{N} : g \in \mathcal{A}_i\})$$

$$\mathcal{F}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha) = Cl^\subseteq_{\mathcal{P}(\mathcal{N})}(\bigcup_{b \in \alpha^{-1}(C^{\prec k})} \{i \in \mathcal{N} : b \in \mathcal{A}_i\})$$

By construction: each set $\mathcal{S}^2_{\langle \mathcal{A}^k_i \rangle}(\alpha)$ is an upset for inclusion; the sets $\langle \mathcal{S}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha) \rangle$ are nested (i.e. $\mathcal{S}^2_{\langle \mathcal{A}^k_i \rangle}(\alpha) \subseteq \cdots \subseteq \mathcal{S}^p_{\langle \mathcal{A}^k_i \rangle}(\alpha)$); each set $\mathcal{F}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha)$ is a lower set for inclusion; and the sets $\langle \mathcal{F}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha) \rangle$ are nested (i.e. $\mathcal{F}^p_{\langle \mathcal{A}^k_i \rangle}(\alpha) \subseteq \cdots \subseteq \mathcal{F}^2_{\langle \mathcal{A}^k_i \rangle}(\alpha)$). Additionally, having disjoint observed sufficient and insufficient coalitions at every exigence level, i.e. $\forall k \in [2.p] \ \mathcal{S}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha) \cap \mathcal{F}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha) = \emptyset$ is a necessary and sufficient condition for the existence of nested sets of coalitions $\langle \mathcal{T}^k \rangle_{k \in [2.p]}$ such that $\forall k \in [2.p], \mathcal{S}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha) \subseteq \mathcal{T}^k \subseteq \mathcal{P}(\mathcal{N}) \setminus \mathcal{F}^k_{\langle \mathcal{A}^k_i \rangle}(\alpha)$.

Theorem 4.2. (*Pairwise formulation of the Non-Compensatory Sorting model*). *An assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$ can be represented in the Non-Compensatory Sorting model if, and only if, there are tuples $\langle \mathcal{A}^k_i \rangle_{i \in \mathcal{N}, \ k \in [2.p]}$ such that:*

1. *(Upset): for each criterion $i \in \mathcal{N}$ and for each exigence level $k \in [2.p]$, $\langle \mathcal{A}^k_i \rangle$ is an upset of $(\mathbb{X}_i, \precsim_i)$; and*
2. *(Nesting): the approved sets are nested according to their exigence level, i.e. for each criterion $i \in \mathcal{N}$, $\mathcal{A}^2_i \subseteq \cdots \subseteq \mathcal{A}^p_i$ (according to a given point of view, an alternative approved at some exigence level k is also approved at any lower exigence level); and*

---

[3] This assumption might also have some relevance w.r.t. intelligibility and parsimony.

3. *(Pairwise separation): for any two exigence levels $k \leq k'$, for each pair of alternatives $(g, b) \in \alpha^{-1}(C^{k'}) \times \alpha^{-1}(C^{k-1})$, there is at least one point of view $i \in \mathcal{N}$ such that $g \in \mathcal{A}_i^{k'}$ and $b \notin \mathcal{A}_i^k$.*

Proof. $[(1, 2, 3) \Rightarrow$ (NCS)]. Given a set of approved sets $\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, k \in [2.p]}$ such that for each exigence level $k \in [2.p]$, $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \precsim_i)$ satisfying conditions 1, 2 and 3, we consider the sets of coalitions $\mathcal{S}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha)$ and $\mathcal{F}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha)$ for each exigence level $k \in [2.p]$. According to the remark just above, α can be represented in the NCS model iff $\mathcal{S}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha) = \emptyset, \forall k \in [2.p]$. Suppose this intersection is not empty for a given $k \in [2.p]$, and let $B \in \mathcal{S}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha)$. By definition of $\mathcal{S}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha)$, there is an exigence level $h \in [k.p]$ and an alternative $g \in \alpha^{-1}(C^h)$ such that $\{i \in \mathcal{N} : g \in \mathcal{A}_i^h\} \subseteq B$. By definition of $\mathcal{F}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha)$, there is an exigence level $h' \in [2.k]$ and an alternative $b \in \alpha^{-1}(C^{h'-1})$ such that $B \subseteq \{i \in \mathcal{N} : b \in \mathcal{A}_i^{h'}\}$. Consequently, there is no criterion $i \in \mathcal{N}$ according to which $g \in \mathcal{A}_i^h$ and $b \notin \mathcal{A}_i^{h'}$, contradicting condition 3. Hence, $\mathcal{S}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i^k \rangle}^k(\alpha) = \emptyset$.

$[\neg(1, 2, 3) \Rightarrow \neg$ (NCS)]. It is obvious that condition 1 and condition 2 are essential to learn an NCS model with nested satisfactory values (enforced by condition 2) and nested sufficient coalitions sets (by construction). Suppose now that condition 1 and condition 2 are satisfied and let $k \in [2.p]$, $k' \in [k.p]$ a pair of exigence levels and (g, b) a pair of alternatives such that $g \in \alpha^{-1}(C^{k'})$, $b \in \alpha^{-1}(C^{k-1})$ and $[(k, k'), (b, g)]$ falsifies condition 3 i.e. $\{i \in \mathcal{N} : g \in \mathcal{A}_i^k\} \subseteq \{i \in \mathcal{N} : b \in \mathcal{A}_i^k\}$. As $g \in \alpha^{-1}(C^{k'})$, the coalition of criteria $\{i \in \mathcal{N} : g \in \mathcal{A}_i^{k'}\}$ is observably sufficient at level k'. As $b \in \alpha^{-1}(C^{k-1})$, the coalition of criteria $\{i \in \mathcal{N} : b \in \mathcal{A}_i^k\}$ is observably insufficient at level k, and even more so at level $k' \geq k$. Hence the intersection $\mathcal{S}_{\langle \mathcal{A}_i^k \rangle}^{k'}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i^k \rangle}^{k'}(\alpha)$ is nonempty, and α cannot be represented in NCS. □

4.3. A SAT formulation for Inv-NCS based on pairwise separation conditions

The Boolean satisfiability formulation for learning an NCS model presented in this section, denoted Φ_α^P, was initially described in Belahcene et al. (2018b) but only focusing on the case with two categories $C^1 \equiv \textsc{Bad} \prec C^2 \equiv \textsc{Good}$. We extend this formulation to the multiple categories case to learn NCS, U^B-NCS and U^C-NCS.

4.3.1. Learning NCS in the case of two categories
The SAT formulation based on pairwise separation initially given in Belahcene et al. (2018b) corresponds to the SAT encoding of both conditions of the Theorem 4.1. First condition which ensures the monotony of scales is represented by a single family of clauses and operates on the same variables as the SAT formulation based on coalitions. In the second condition, additional binary variables are defined in order to represent the separation between the alternatives. A unique family of logical clauses represent the separation concept of the theorem and additional clauses and binary variables are required in order to express this representation in SAT language. *Encoding* Similarly to the formulation Φ_α^C described in Section 4.1, the formulation Φ_α^P operates on two types of variables:

- '*a*' variables, representing the approved sets, with the exact same semantics as their counterpart in Φ_α^C, i.e.

$$a_{i,x} = \begin{cases} 1, & \text{if } x \in \mathcal{A}_i \text{ i.e. } x \text{ is approved according to } i; \\ 0, & \text{else.} \end{cases}$$

- auxiliary '*s*' variables, indexed by a criterion $i \in \mathcal{N}$, an alternative g assigned to \textsc{Good} and an alternative b assigned to \textsc{Bad}, assessing if the alternative g is positively separated from b according to the criterion i, i.e.

$$s_{i,g,b} = \begin{cases} 1, & \text{if } g \in \mathcal{A}_i \text{ and } b \notin \mathcal{A}_i; \\ 0, & \text{else.} \end{cases}$$

Φ_α^P is the conjunction of four types of clauses: ϕ_α^{P1} ensuring each $\mathcal{A}_i$ is an upset, ϕ_α^{P2} ensuring $[s_{i,g,b} = 1] \Rightarrow [g \in \mathcal{A}_i]$, ϕ_α^{P3} ensuring $[s_{i,g,b} = 1] \Rightarrow [b \notin \mathcal{A}_i]$, and ϕ_α^{P4} ensuring each pair (g, b) is positively separated according to at least one criterion.

Definition 4.3. Given an instance of Inv-NCS with two categories and an assignment $\alpha : \mathbb{X}^\star \to \{ \textsc{Bad} \prec \textsc{Good} \}$, we define the Boolean function Φ_α^P with variables $\langle a_{i,x} \rangle_{i \in \mathcal{N}, x \in \mathbb{X}^\star}$ and $\langle s_{i,g,b} \rangle_{i \in \mathcal{N}, g \in \alpha^{-1}(Good), b \in \alpha^{-1}(bad)}$, as the conjunction of clauses:

$$\phi_\alpha^P = \phi_\alpha^{P1} \wedge \phi_\alpha^{P2} \wedge \phi_\alpha^{P3} \wedge \phi_\alpha^{P4}$$

$$\phi_\alpha^{P1} = \bigwedge_{i \in \mathcal{N}} \bigwedge_{x' \succsim_i x \in \mathcal{X}^\star} (a_{i,x'} \vee \neg a_{i,x})$$

$$\phi_\alpha^{P2} = \bigwedge_{i \in \mathcal{N}, g \in \alpha^{-1}(Good), b \in \alpha^{-1}(Bad)} (\neg s_{i,g,b} \vee \neg a_{i,b})$$

$$\phi_\alpha^{P3} = \bigwedge_{i \in \mathcal{N}, g \in \alpha^{-1}(Good), b \in \alpha^{-1}(Bad)} (\neg s_{i,g,b} \vee a_{i,g})$$

$$\phi_\alpha^{P4} = \bigwedge_{g \in \alpha^{-1}(Good), b \in \alpha^{-1}(Bad)} (\bigvee_{i \in \mathcal{N}} s_{i,g,b})$$

The formulation is compact: $O(|\mathcal{N}|.|\mathbb{X}|^2)$ variables, $O(|\mathcal{N}|.|\mathbb{X}|^2)$ binary clauses and $O(|\mathbb{X}|^2)$ $|\mathcal{N}|$-ary clauses, whereas the number of '*t*' variables in the first formulation increases exponentially with the number of criteria.

Should ϕ_α^P be satisfiable, the set $\mathcal{T}$ of sufficient coalitions is not uniquely identified by the values of '*a*' and '*s*' variables of one of its models. Indeed, if $\langle a_{i,x} \rangle, \langle s_{i,g,b} \rangle$ is an antecedent of 1 by ϕ_α^P, then the parameter $\omega = (\langle \mathcal{A}_i \rangle, \mathcal{T})$ with accepted sets defined by $\mathcal{A}_i = \{x \in \mathbb{X} : a_{i,x} = 1\}$ and any upset $\mathcal{T}$ of $(\mathcal{P}(\mathcal{N}), \subseteq)$ of sufficient coalitions containing the upset $\mathcal{S}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and disjoint from the lower set $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is a solution of this instance. Therefore, among the sets of sufficient coalitions compatible with the values of '*a*' and '*s*' variables, we can identify two specific ones, $\mathcal{T}_{\max}$ and $\mathcal{T}_{\min}$. We will also denote by $\mathcal{T}_{rand}$, a randomly chosen compatible set of sufficient coalitions.

4.3.2. Learning NCS with more than two categories
When there are more than two categories, the sets of variables and clauses need to be extended in order to characterize the NCS model.

- '*a*' variables are also indexed by an exigence level $k \in [2.p]$, i.e.

$$a_{i,k,x} = \begin{cases} 1, & \text{if } x \in \mathcal{A}_i^k \text{ i.e. } x \text{ is approved according to } i \\ & \text{at exigence level } k; \\ 0, & \text{else.} \end{cases}$$

- '*s*' variables are also indexed by a *pair* of exigence levels $(k, k') \in [2.p]^2, k \leq k'$, with $g \in \alpha^{-1}(C^{k'})$, $b \in \alpha^{-1}(C^{k-1})$, so that

$$s_{i,k,k'g,b} = \begin{cases} 1, & \text{if } g \in \mathcal{A}_i^{k'} \text{ and } b \notin \mathcal{A}_i^k; \\ 0, & \text{else.} \end{cases}$$

These additional indices do not refer to new variables, but allow to tie the s variables representing pairwise separation to the a variables representing acceptance at the proper exigence level.

As it was introduced in Belahcene et al. (2018b), in the second formulation we learn the nested approved sets $\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, k \in [2.p]}$

with which we identify the nested sets of sufficient coalitions $\langle \mathcal{S}^k_{\langle \mathcal{A}_i \rangle}(\alpha) \rangle$ and insufficient coalitions $\langle \mathcal{F}^k_{\langle \mathcal{A}_i \rangle}(\alpha) \rangle$. Approved sets are constrained so that the intersection between the sets of observably sufficient and insufficient coalitions is empty. Leveraging Theorem 4.2, this ensures that the reference assignments are fully restored.

Definition 4.4. Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$, we define the Boolean function $\Phi^{P'}_\alpha$ with variables $\langle a_{i,k,x} \rangle_{i \in \mathcal{N},\ k \in [2.p], x \in \mathbb{X}^\star}$ and

$\langle s_{i,k,k',g,b} \rangle_{i \in \mathcal{N},\ k \in \{2.p\},\ k' \in \{k.p\},\ g \in \alpha(C^{k'}),\ b \in \alpha(C^{k-1})}$, as the conjunction of clauses:

$$\Phi^{P'}_\alpha = \phi^{P'1} \wedge \phi^{P'2} \wedge \phi^{P'3} \wedge \phi^{P'4} \wedge \phi^{P'5}$$

$$\phi^{P'1}_\alpha = \bigwedge_{i \in \mathcal{N},\ k \in [2.p]} \bigwedge_{x' \succeq_i x \in \mathbb{X}^\star} (a_{i,k,x'} \vee \neg a_{i,k,x})$$

$$\phi^{P'2}_\alpha = \bigwedge_{i \in \mathcal{N},\ k < k' \in [2.p],\ x \in \mathbb{X}^\star} (a_{i,k,x} \vee \neg a_{i,k',x})$$

$$\phi^{P'3}_\alpha = \bigwedge_{i \in \mathcal{N},\ k \in [2.p],\ k' \in [k.p]} \bigwedge_{g \in \alpha^{-1}(C^{k'}),\ b \in \alpha^{-1}(C^{k-1})} (\neg s_{i,k,k',g,b} \vee \neg a_{i,k,b})$$

$$\phi^{P'4}_\alpha = \bigwedge_{i \in \mathcal{N},\ k \in [2.p],\ k' \in [k.p]} \bigwedge_{g \in \alpha^{-1}(C^{k'}),\ b \in \alpha^{-1}(C^{k-1})} (\neg s_{i,k,k',g,b} \vee a_{i,k',g})$$

$$\phi^{P'5}_\alpha = \bigwedge_{k \in [2.p],\ k' \in [k.p]} \bigwedge_{g \in \alpha^{-1}(C^{k'}),\ b \in \alpha^{-1}(C^{k-1})} (\bigvee_{i \in \mathcal{N}} s_{i,k,k',g,b})$$

The remarks made about an efficient implementation of Φ^C_α are still valid: many clauses are redundant in $\phi^{P'1}_\alpha$ and $\phi^{P'2}_\alpha$ and can safely be ignored. The remark concerning the non-uniqueness of $\mathcal{T}$ in the case of two categories also applies for more than two categories to $\mathcal{T}^k$ which are not uniquely defined by $\Phi^{P'}_\alpha$.

Corollary 4.1. *Given a context, an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$ can be represented in the Non-Compensatory Sorting model if, and only if $\Phi^{P'}_{\alpha, NCS}$ is satisfiable.*

A specific analysis of how to extend Definition 4.3 to more than two categories when learning a U^B-NCS or a U^C-NCS model is detailed in Appendix A.

5. MaxSAT relaxations for Inv-NCS

The previous section introduced mathematical and computational tools addressing the *decision* problem: can a given assignment be represented in the Non-Compensatory Sorting model (or one of its variants)? This set of tools has an important theoretical significance, and can also serve as a base for practical applications– see e.g. Belahcene et al. (2018b) for an application in an *accountability* setting, where the representation theorem (Theorem 4.1) is leveraged to provide procedural regularity certificates with good properties in terms of computational hardness and privacy preservation, or jurisprudential explanations, should the outcome of the sorting process be contested. Nevertheless, this approach is not suited to the problem of learning a suitable NCS model from real data, because it does not tolerate the presence of noise in the data. There are numerous reasons for the input data not to reflect perfectly the model, e.g.: imperfections in the assessment of performance according to some point of view; mistaken assignment of an alternative to a category; or simply the oversimplification of reality represented by the model.

In this section, we address this issue by providing a relaxation of the decision formulations: instead of finding an NCS model restoring all examples of the learning set (or, probably, die trying), we try to find the model that restores the most. This approach is similar to the *empirical risk minimization* approach that is central in Machine Learning for supervised classification problems, using the 0–1 loss. While it is a common practice in ML to use a convex surrogate of the 0–1 loss to immensely speed up the learning

process, we embrace the computationally much more demanding exact approach, because we believe the benefits are high in terms of accountability–we are absolutely sure no one can challenge the output model on the basis of a better restoration of the learning set–while the computational cost can be kept low enough–because the number of criteria and of learning examples are often low in typical applications, and because we propose a computationally efficient approach.

We formulate the relaxed *optimization* problem of finding the subset of learning examples (reference alternatives together with their assignment) correctly restored of maximum cardinality with a *soft constraint* approach, using the language of weighted MaxSAT. This framework, derived from the SAT framework, is based on a conjunction of clauses $\bigwedge c_i$ where each clause c_i is given a non-negative weight w_i, and maximizes the total weight of the satisfied clauses. In order to translate exactly our problem in this language, we leverage two basic techniques: we introduce switch variables 'z' allowing to precisely monitor the soft clauses we are ready to see violated, as opposed to hard clauses that remain mandatory; and we use big-stepped tuples of weights $w_1, \ldots, w_k$ with $w_1 \gg \cdots \gg w_k$ allowing to specify lexicographically ordered goals in an additive framework.

5.1. A MaxSAT relaxation for Inv-NCS based on coalitions

This section elaborates on the SAT formulation introduced in Section 4.1. The MaxSAT extension of the formulation obtained when following a strategy based on the explicit representation of coalitions of criteria is based on the specification of the reference alternative to relax in order to remove conflicts in the clauses. For this purpose, we define the following additional binary variables:

- 'z' variables, indexed by an alternative x, represent the set of alternatives properly classified by the inferred model, with the following semantic: $z_x = 1 \Leftrightarrow \alpha^{-1}(x) = NCS_\omega(x)$ i.e. the alternative x is properly classified

These variables are introduced in some clauses to serve as switches:

- For any exigence level $k \in [2.p]$, let $B \subseteq \mathcal{N}$ a coalition of criteria, and x an alternative assigned to C^{k-1} by α. If $z_k = 1$ and $B \subseteq \{i \in \mathcal{N} : x \in \mathcal{A}^k_i\}$ then $t_{B,k} = 0$. This leads to the following conjunction of clauses:

$$\phi^{\widetilde{C5}}_\alpha = \bigwedge_{B \subseteq \mathcal{N},\ k \in [2.p]} \bigwedge_{x \in \alpha^{-1}(C^{k-1})} (\bigvee_{i \in B} \neg a_{i,k,x} \vee \neg t_{B,k} \vee \neg z_x)$$

- For any exigence level $k \in [2.p]$, let $B \subseteq \mathcal{N}$ a coalition of criteria, and x an alternative assigned to C^k by α. If $z_k = 1$ and $B \subseteq \{i \in \mathcal{N} : x \in \mathcal{A}^k_i\}$ then $t_{\mathcal{N} \setminus B,k} = 0$. This leads to the following conjunction of clauses:

$$\phi^{\widetilde{C6}}_\alpha = \bigwedge_{B \subseteq \mathcal{N},\ k \in [2.p]} \bigwedge_{x \in \alpha^{-1}(C^k)} (\bigvee_{i \in B} a_{i,k,x} \vee t_{\mathcal{N} \setminus B,k} \vee \neg z_x)$$

The objective in the MaxSAT formulation is to maximize the portion of alternatives properly classified, this is the subject of the following soft clause:

$$\phi^{goal}_\alpha = \bigwedge_{x \in \mathbb{X}^\star} z_x \tag{8}$$

The MaxSAT extension of the first formulation is the conjunction of the first four clauses of the SAT formulation given in Definition 4.1 and clauses $\phi^{\widetilde{C5}}_\alpha$, $\phi^{\widetilde{C6}}_\alpha$ and ϕ^{goal}_α.

Clauses composing the conjunctions ϕ^{C1}_α, ϕ^{C2}_α, ϕ^{C3}_α, ϕ^{C4}_α, $\phi^{\widetilde{C5}}_\alpha$ and $\phi^{\widetilde{C6}}_\alpha$ are hard, associated to the weight w_{max}, and we associate to ϕ^{goal}_α the weight w_1 such that $w_{max} > |\mathbb{X}^\star| w_1$.

Model variants Same modifications as in the SAT formulation are required to learn U^B-NCS and U^C-NCS models with noisy preference information:
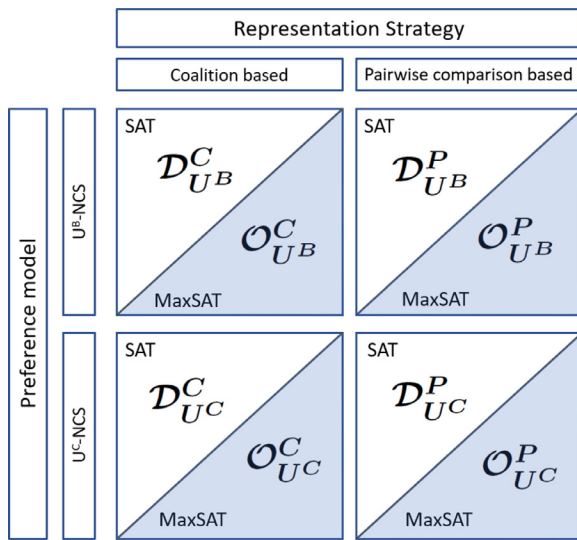
Fig. 4. The eight approaches considered.

- U^B-NCS (Unique profiles): Drop the index k concerning the exigence level for the 'a' variables, ignore the conjunction over exigence levels for clauses ϕ_α^{C1}, and ignore clauses ϕ_α^{C2} altogether;
- U^C-NCS (Unique set of sufficient coalitions): Drop the index k concerning the exigence level for the 't' variables, ignore the conjunction over exigence levels for clauses ϕ_α^{C3} and ignore clauses ϕ_α^{C4} altogether.

5.2. A MaxSAT relaxation for Inv-NCS based on pairwise separation conditions

This section elaborates on the SAT formulation introduced in Section 4.3, following a representation strategy based on the pairwise separation of alternatives.

In the case of two categories, switch variables 'z' have the same indexation and semantics as in the previous section. They are introduced in the clauses representing the pairwise separation constraints:

$$\phi_\alpha^{\widetilde{P4}} = \bigwedge_{g\in\alpha^{-1}(Good),\ b\in\alpha^{-1}(Bad)}\left(\bigvee_{i\in\mathcal{N}} s_{i,g,b} \vee \neg z_b \vee \neg z_g\right)$$

They also appear in the clause ϕ_α^{goal} (see Eq. (8)) formulating our objective of restoring as many learning examples as we can.

The weighted MaxSAT relaxation of the SAT formulation obtained following the representation strategy based on pairwise separation of alternatives, in the case of two categories, is the conjunction of clauses $\phi_\alpha^{P1} \wedge \phi_\alpha^{P2} \wedge \phi_\alpha^{P3} \wedge \phi_\alpha^{\widetilde{P4}}$, where each clause is hard and receives the weight $w_{\max}$, and of the clause ϕ_α^{goal} with weight w_1 such that $w_{\max} > |\mathbb{X}^\star|w_1$.

The generalizations of this MaxSAT formulation to the case with multiple categories, including adaptations geared towards learning U^B-NCS and U^C-NCS variants of the Non-Compensatory Sorting model, are provided in Appendix B.

6. Computational study

In this section, we present an empirical study that evaluates the intrinsic and comparative performances of the approaches presented in Sections 4 and 5. There are eight of them, depicted on Fig. 4 and specified by three binary parameters:

- the Non-Compensatory Sorting model of preference sought, either with a *unique* boundary *profile* (subscript **U^P**), or with a *unique* set of sufficient *coalitions* (subscript **U^C**);

- the representation strategy adopted, based either on the explicit representation of the *coalitions* of criteria (superscript **C**) or on the *pairwise* separation of alternatives (superscript **P**); and
- the problem description, either *deciding* whether an instance can be represented in the model ($\mathcal{D}$) with a SAT solver, or *optimizing* the ability of the model to represent the assignment ($\mathcal{O}$) with a MaxSAT solver.

Note that the performances of $\mathcal{D}_{U^C}^C$ for learning U^C (Section 4.1) have already been proved to be superior to MIP approaches by Belahcene et al. (2018a).

6.1. Experimental design

The experimental plan consists of generating random instances of the Inv-NCS problem, applying one of the eight approaches described above, and measuring several performance indicators. We detail the instance generator, the implementation of the approaches and the indicators in the following sections.

6.1.1. Instance generation

Each instance consists of a set of alternatives $\mathbb{X}^*$ (described by tuples of evaluations on a set of criteria $\mathcal{N}$), a set of categories $C^1 \prec \ldots \prec C^p$, and the assignment of the former to the latter. We set the number of categories p to three. The set of alternatives is governed by two parameters –the number of criteria $|\mathcal{N}|$ and the number of reference alternatives $|\mathbb{X}^\star|$ – that we consider exogenous and we try to assess their respective influence on the performance indicators. Note that this design is similar to a supervised classification context, where $|\mathbb{X}^\star|$ and $|\mathcal{N}|$ are respectively the number of rows and columns of the dataset. Instances are sampled uniformly from the cube $[0,1]^{|\mathcal{N}|}$: we have considered the least favourable case where all the criteria take continuous values.

The assignment of alternatives to categories depends on the type of model sought and the problem description. In order to ensure that preference data represents a real decision problem, we use a decision model to generate it, and, in particular, a model compatible to the Non-Compensatory stance we are postulating:

- In the case of U^C-NCS, we use an MR-Sort model for generating the learning set, a model that particularizes U^C by postulating the set of sufficient coalitions has an additive structure (see Section 2). It is randomly generated using the following procedure: a set of limit profiles $\langle b\rangle$ is generated by uniformly sampling $p-1$ numbers in the interval $[0,1]$ and sorting them in ascending order, for all criteria; the voting powers $\langle w\rangle$ are generated by sampling $|\mathcal{N}|-1$ numbers in the interval $[0,1]$, sorted and used as the cumulative sum of weights; the majority threshold λ is sampled with uniform probability in the interval $[0.5,1]$.
- In the case of U^B-NCS we use a model with a unique profile and nested sets of sufficient coalitions of criteria at each exigence level, each with an additive structure, i.e., weights attached to criteria and and a majority threshold. It is randomly generated using the following procedure: a single profile b is generated by uniformly sampling a tuple in $[0,1]^{\mathcal{N}}$; the voting weights $\langle w\rangle$ are generated by sampling $|\mathcal{N}|-1$ numbers in the interval $[0,1]$, sorted and used as the cumulative sum of weights; the majority thresholds $\langle \lambda\rangle$ are then randomly chosen by sampling $p-1$ numbers with uniform probability in the interval $[0.5,1]$ and sorting them in ascending order.

Once the *ground truth* model is generated, which is by design compatible to the hypothesis class we are working with, we consider two ways of assigning alternatives to categories, depending on the problem formulation we are considering.

- For decision approaches, we directly assign the alternatives to categories according to the ground truth. Therefore, these approaches should always succeed in finding the parameters of a model extending the reference assignment.
- For optimization approaches, we introduce a proportion μ of assignment errors in the learning set. The assignment of a subset of reference alternatives is randomly replaced, with uniform probability, by the successor or predecessor category compared to the ground truth assignment.

6.1.2. Solving the instances

This experimental study is run on a laptop with Windows 10 (64 bit) equipped with an Intel(R) Xeon(R) CPU E5-1620 v4 @3.5GHz and 32 GB of RAM.

For decision approaches, we translate the assignment into a Boolean satisfaction problem, described by sets of variables and clauses, for both representation strategies and both preference models, as described in Section 4. The SAT instances are written in a file in DIMACS format, and passed to a command line SAT solver - CryptoMiniSat 5.0.1.

For optimization approaches, we translate the assignment into a Boolean satisfaction problem, described by sets of variables and clauses and an objective function, for both representation strategies and both preference models, as described in Section 5. The MaxSAT instances are passed to a command line MaxSAT solver QmaxSAT in the required format.

When using the representation strategy based on the explicit representation of the set of coalitions of criteria, each solution of the SAT/MaxSAT problem found by the solver can directly be interpreted in terms of parameters of an NCS model (either of the U^B or the U^C subtype). This is not exactly the case with the representation strategy based on pairwise separation of alternatives: the SAT/MaxSAT solution explicitly describes the approved sets of value on each criterion and at each exigence level (i.e. the boundary profiles), but the sets of sufficient coalitions are left implicit, and are solely described in terms of an upper and a lower bound. In the context of this experimental study, we are interested in resolute and precise decision models – hence it is necessary to complete this irresolute (or imprecise) strategy with a second strategy for picking a specific (nesting of) upset(s) of sufficient coalition inside the band of possible sets. We consider three such post-processing strategies: i) $\mathcal{T} = \mathcal{T}_{\min}$, systematically returning the lower bound, ii) $\mathcal{T} = \mathcal{T}_{rand}$, returning a random nesting of upsets satisfying the constraints; and iii) $\mathcal{T} = \mathcal{T}_{\max}$, returning systematically the upper bound.

6.1.3. Performance indicators

The performance indicators of interest are the computing time, the restoration rate (the proportion of the learning set correctly represented by the output model), and the generalization index measuring the alignment between the output model with the ground truth.

So as to monitor the learning process, we control the level of noise in the input data through the parameter μ, and we measure the proportion of reference assignments that are correctly restored by the learning process. This *restoration rate* should be equal to one in the case of approaches addressing the decision problem (as there is no noise), and at least equal to $1 - \mu$ for approaches addressing the optimization problem.

The computing performance is measured in practice, by solving actual instances of the problem and reporting the *computation time* required by the solver.

In order to appreciate how "close" a computed model to the ground truth from which the assignment examples were generated, and thus to monitor potential overfitting, we proceed as follows: we sample a large set of n profiles in $\mathbb{X} = [0, 1]^{\mathcal{N}}$ and com-

Table 5
Computation time in the baseline configuration (128 ref. alternatives, 9 crit., 3 categ.) to learn a U^B model.

	$\mathcal{D}^C_{U^B}$	$\mathcal{D}^P_{U^B}$
Max	0.169s	0.293s
2nd quartile	0.141s	0.184s
Median	0.126s	0.148s
1st quartile	0.118s	0.111s
Min	0.108s	0.06s

Table 6
Computation time in the baseline configuration (128 ref. alternatives, 9 crit., 3 categ.) to learn a U^C model.

	$\mathcal{D}^C_{U^C}$	$\mathcal{D}^P_{U^C}$
Max	0.161s	0.584s
2nd quartile	0.139s	0.389s
Median	0.131s	0.337s
1st quartile	0.123s	0.256s
Min	0.104s	0.097s

pute the assignment of these profiles according to the original and computed models. On this basis, we compute the *generalization index*: the proportion of "correct" assignments, i.e. profiles which are assigned to the same category by the ground truth and the inferred model.

6.2. Model retrieval with decision approaches

In this section, we study the behavior of the decision approaches, when fed with synthetic data matching the hypothesis (i.e. either coming from a specific U^B or U^C NCS model). More particularly, we monitor the restoration rate (which is expected to reach 100%), the computation time and the generalization index when applying each strategy (and, concerning the one based on the pairwise separation of alternatives, of three specific post-processing strategies concerning the choice of the nested sufficient coalitions), i.e. for the approaches $\mathcal{D}^C_{U^B}$, $\mathcal{D}^C_{U^C}$, $\mathcal{D}^P_{U^B}$ and $\mathcal{D}^P_{U^C}$, as functions of the number of reference alternatives $|\mathbb{X}^\star|$ and the number of criteria $|\mathcal{N}|$.

We explore a specific subset of the parameter space: we consider a baseline configuration, with 3 categories, 9 criteria and 128 reference alternatives, and we consider the configurations deviating from the baseline on a single parameter – either $|\mathbb{X}^\star| = 128$ and $|\mathcal{N}| \in \{3, 5, 7, 11\}$, or $|\mathbb{X}^\star| \in \{32, 64, 256, 512, 1024\}$ and $|\mathcal{N}| = 9$. For each configuration and for both models U^C and U^B, we sample 50 instances, then solve each of them according to both strategies.

6.2.1. Restoration rate

As expected, the restoration rate, for every model and strategy, is uniformly equal to one.

6.2.2. Computing time

For each NCS model (U^B and U^C), for each strategy under scrutiny (coalition based, and pairwise separation based), and for the set of considered parameters governing the input, the computation time ranges from below the tenth of a second to some dozens of minutes. Table 5 (respectively Table 6) depicts the distribution of the computation time for the baseline situation (128 reference assignments, 9 criteria and 3 categories) of implementing each strategy to learn a U^B model (resp. a U^C model). In this configuration, the strategy based on coalitions ($\mathcal{D}^C$) is slightly faster than the one based on pairwise separation ($\mathcal{D}^P$) when learning a U^B model and three times faster when learning a U^C model. The
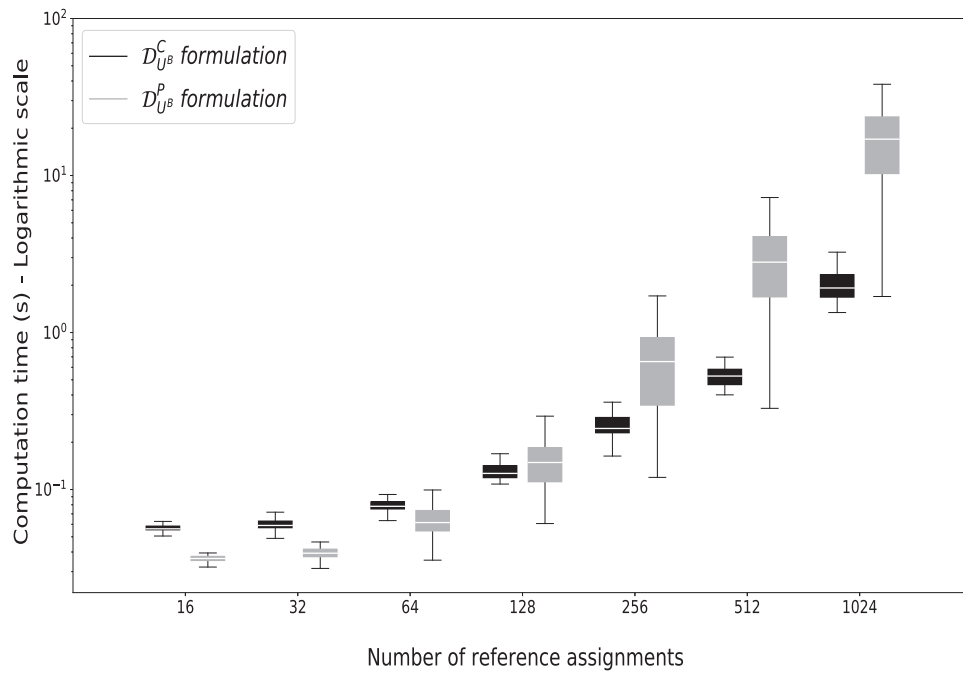
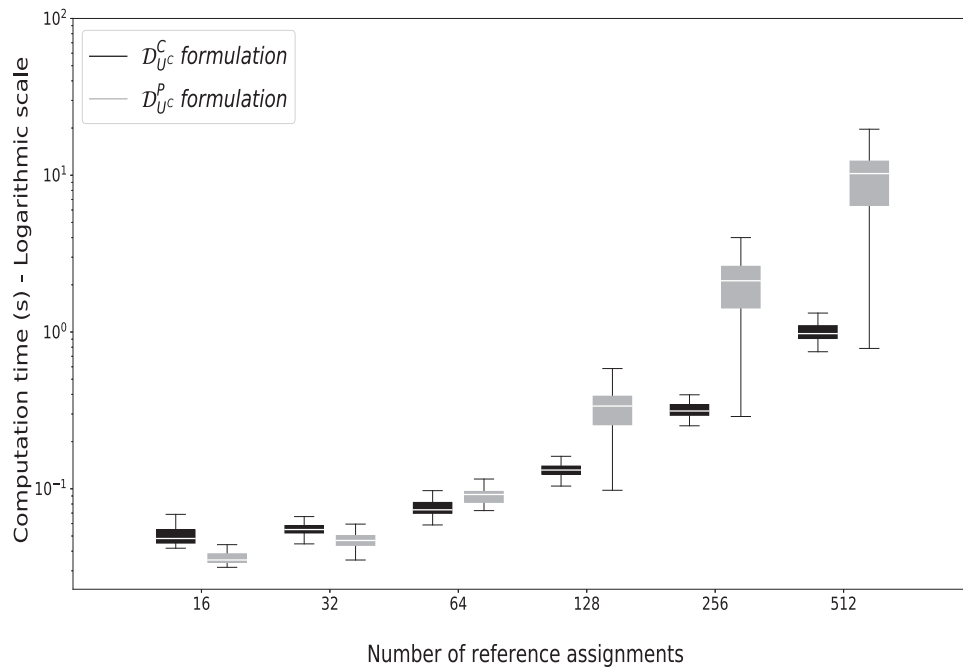**Fig. 5.** Computation time by number of ref. assignments (9 crit., 3 categ.) to learn a U^B model.

Fig. 6. Computation time by number of ref. assignments (9 crit., 3 categ.) to learn a U^C model.

distribution of the computing time of each formulation is very tight around its center.

In order to assess the influence of the parameters governing the size and complexity of the input, we explore situations differing from the baseline on a single parameter:

- The number of reference assignments $|\mathbb{X}^\star|$: Figs. 5 and 6 indicate that the distribution of the computing time for both strategies and for both U^B and U^C models remains tightly grouped around its central value. It also shows that this value steadily increases with the number of reference assignments. For both strategies, the log-log plots are all consistent with a linear dependency between $\log t$ and $\log |\mathbb{X}^\star|$, indicating the soundness

of power law $t \propto |\mathbb{X}^\star|^\beta$. The observed slopes are consistent with $\beta_C = 1$ (i.e. $t \propto |\mathbb{X}^\star|$) for the representation strategy based on coalitions, and $\beta_P = 2$ (i.e $t \propto |\mathbb{X}^\star|^2$) for the representation strategy based on the pairwise separation of alternatives.

- The number of criteria $|\mathcal{N}|$: Figs. 7 and 8 indicate for each NCS variants, the distribution of the computing time for both strategies. It can be observed that these series remain tightly grouped around their central value and this value steadily increases with the number of criteria. These observations are consistent with the hypotheses $t \propto |\mathcal{N}|$ for the representation strategy based on the pairwise separation of alternatives, and $t \propto 2^{|\mathcal{N}|}$ for the strategy based on coalitions of criteria.
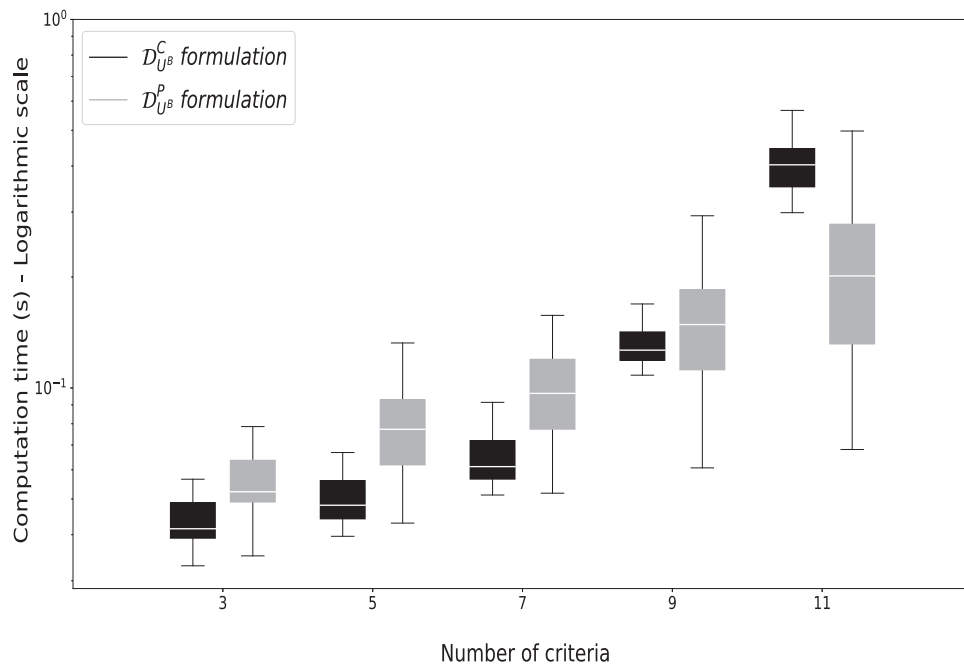
Fig. 7. Computation time by number of criteria (128 ref. assignments and 3 categ.) to learn a U^B model.
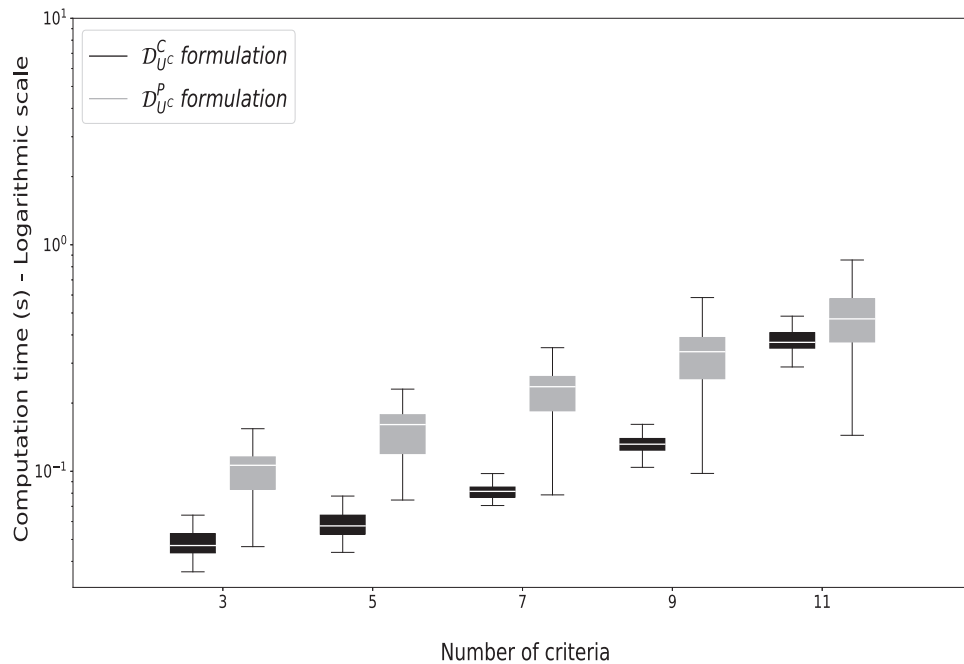
Fig. 8. Computation time by number of criteria (128 ref. assignments and 3 categ.) to learn a U^C model.

6.2.3. Results on the ability of the inferred model to restore the original one

When applied to learn both NCS variants (U^B and U^C), the strategy based on pairwise separation returns an acceptable nesting of upset of sufficient coalitions, defined by lower and upper bounds. This strategy needs to be completed by a post-processing phase dedicated to pinpoint a single nesting of upsets. While this phase has no bearing on the restoration rate, and takes negligible time, it has a measurable impact on the generalization index.

To identify the upset that best restores the simulated sorting model ($1-U^B$ and MR-Sort), we study the three following post-processing strategies: $\mathcal{T} = \mathcal{T}_{\min}$, $\mathcal{T} = \mathcal{T}_{rand}$ and $\mathcal{T} = \mathcal{T}_{\max}$. T-Student

tests ($\alpha = 5\%$) show that for U^B and U^C the generalization index when $\mathcal{T} = \mathcal{T}_{\min}$ is always at least as good as the other two variants regardless the number of criteria, alternatives (and even categories for $p \in \{2, 3, 4, 5\}$); see for instance the baseline configuration Table 7. Consequently, for ease of presentation, we only plot results concerning the post-processing strategy $\mathcal{T} = \mathcal{T}_{\min}$.

The first two columns of Table 7 depicts the distribution of the proportion of correct assignments (as compared to the ground truth) for the baseline situation (128 reference assignments, 9 criteria and 3 categories). T-Student test ($\alpha = 5\%$) shows that the difference between the two distributions is not significant.

Fig. 9. Generalization index by number of reference assignments (9 criteria and 3 categories) to learn a U^B model.

Table 7
Generalization index for both SAT formulations in the baseline configuration (128 reference assignments, 9 criteria and 3 categories) to learn a U^B model.

	$\mathcal{D}_{U^B}^C$	$\mathcal{D}_{U^B}^P$		
		$\mathcal{T} = \mathcal{T}_{min}$	$\mathcal{T} = \mathcal{T}_{rand}$	$\mathcal{T} = \mathcal{T}_{max}$
Max	96.4%	98%	97%	97%
2nd quartile	92.3%	91.4%	89%	89%
Median	89%	90%	85.7%	85.7%
1st quartile	83.4%	86.6%	80.8%	80.8%
Min	75.4%	79.8%	73%	73%

Figs. 9 and 10 represent the variations of the alignment of the models yielded by both formulations with the ground truth with respect to the problem settings when learning a U^B model (respectively U^C model) and applying each strategy. The experimental results display a tendency towards a degradation of this alignment as the number of criteria increases. Conversely, as expected, increasing the number of reference assignments noticeably enhances the generalization index, up to 100%. The implementations of both strategies seem to behave in a similar manner with respect to the variations of these parameters.

6.3. Tolerance for error with optimization approaches

In this section, we study the behavior of the optimization approaches, when fed with synthetic data that deviate from the model hypothesis (i.e. either coming from a specific U^B or U^C NCS model) in a controlled manner, through the incorporation of a proportion μ of noise. More particularly, we monitor the restoration rate (which is expected to reach at least $1 - \mu$), the computation time and the generalization index, when applying each strategy, i.e. for approaches $\mathcal{O}_{U^B}^C, \mathcal{O}_{U^C}^C, \mathcal{O}_{U^B}^P$ and $\mathcal{O}_{U^C}^P$, as functions of the number of reference alternatives $|\mathbb{X}^*|$, the number of criteria $|\mathcal{N}|$, and the noise rate μ.

In this paper, the notion of noise on the learning set is defined as a misclassification of an alternative, i.e., an error from the decision maker in the choice of the category. More precisely, the assignment of a subset of reference alternatives is randomly replaced, with uniform probability, by the successor or predecessor category compared to the ground truth assignment. This is the way we have implemented the noise in our experiment.[4]

We explore a specific subset of the parameter space: we consider a baseline configuration, with 3 categories, 5 criteria, 128 reference alternatives and 10% noise rate, and we consider the configurations deviating from the baseline on a single parameter – $|\mathbb{X}^*| = 128$, $|\mathcal{N}| \in \{3, 7, 9, 11\}$ and $\mu = 0.1$; or $|\mathbb{X}^*| \in \{32, 64, 256\}$, $|\mathcal{N}| = 5$ and $\mu = 0.1$; or $|\mathbb{X}^*| = 128$, $|\mathcal{N}| = 5$ and $\mu \in \{0.05, 0.15, 0.2\}$. For each configuration and for both models U^C and U^B, we sample 50 instances, then solve each of them according to both strategies.

6.3.1. Restoration rate

Plotting the restoration rate allows to monitor the learning process. The experimental results show that, when learning a given subtype of NCS model (either U^B or U^C), the models learned by implementing both strategies (either based on coalition or pairwise separation) reproduce the same portion of the learning set and at least $(1 - \mu) * |\mathbb{X}^*|$ assignment examples. This is some experimental evidence of the validity of the MaxSAT formulations stemming from both representation strategy.

The results display a tendency towards a degradation of the restoration rate distribution as the number of alternatives or the noise rate increases. Conversely, increasing the number of criteria noticeably enhances the restoration rate.

---

[4] Note that there exist alternative ways to consider noisy expression of preferences. On of these is to consider that the errors in the learning set is related to the values/performances of alternatives in the learning set. Such noise is indeed relevant in applications where the learning set correspond to historical data in which performances of examples can be erroneous.
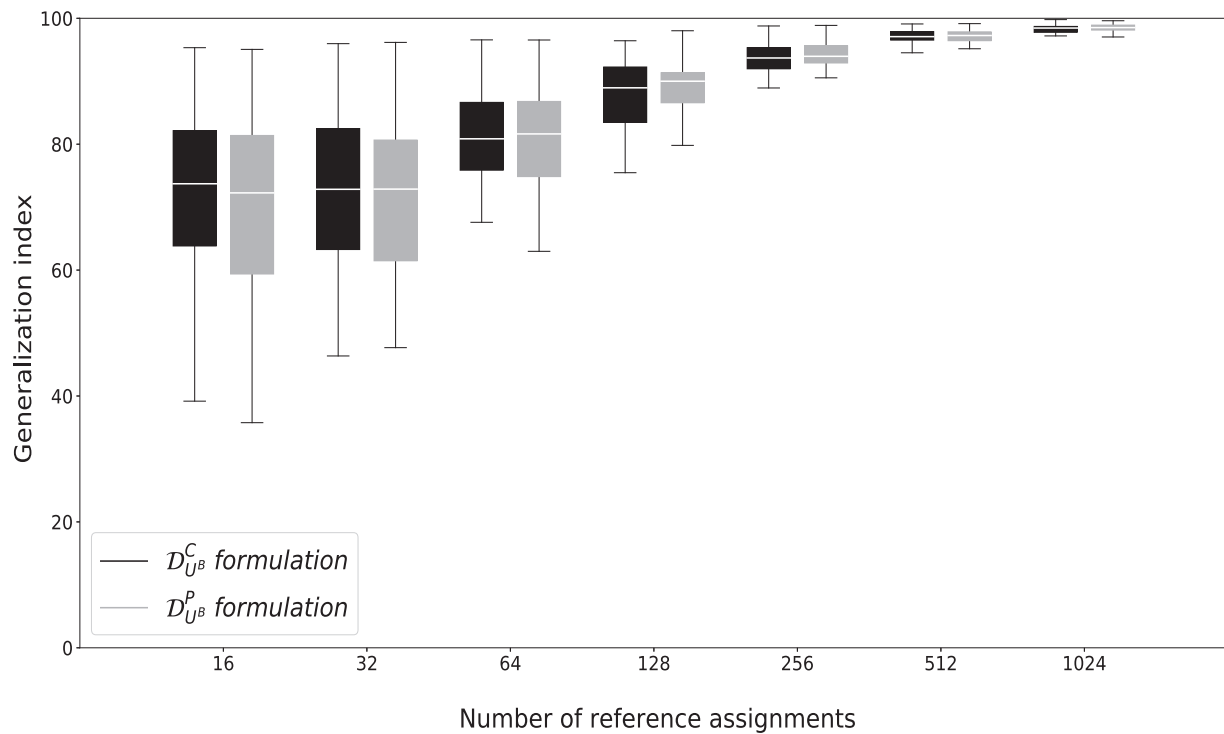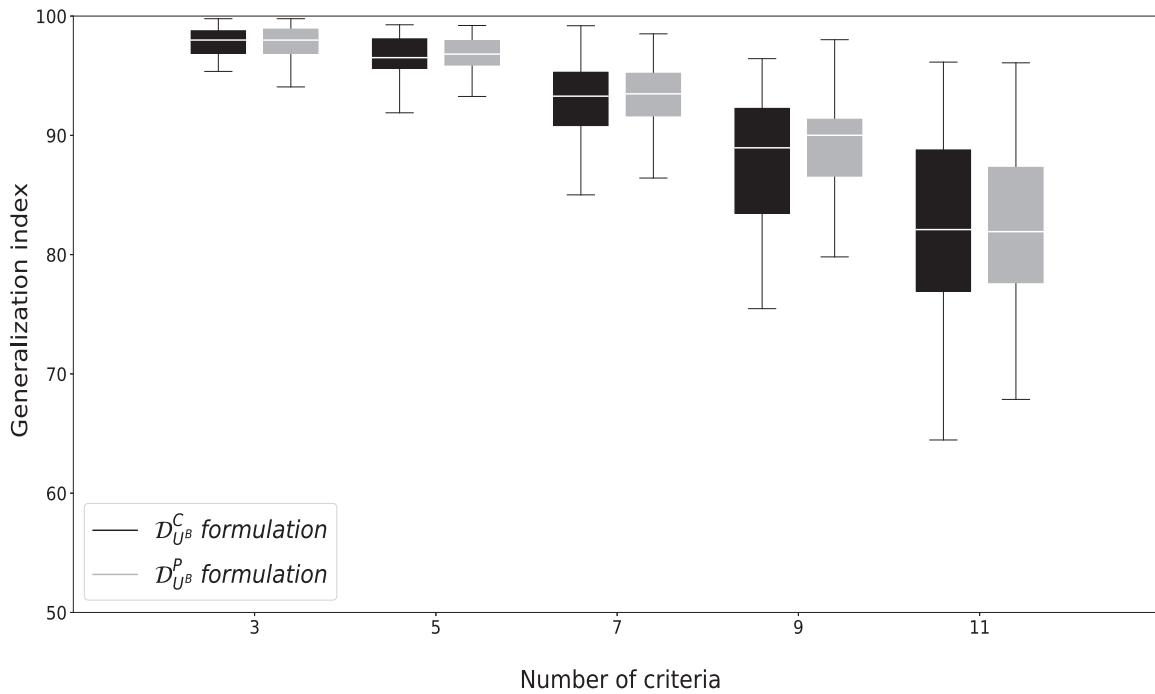
Fig. 10. Generalization index by number of criteria (128 reference assignments and 3 categories) to learn a U^B model.

Fig. 11. Restoration rate by number of reference assignments (5 criteria, 3 categories and 10% noise) to learn a U^B model.

- The number of reference assignments $|\mathbb{X}^\star|$: when the number of learning points (Figs. 11 for U^B and 12 for U^C), we observe a convergence of the restoration rate towards its lower bound $(1 - \mu) * |\mathbb{X}^\star|\%$ (in this case 0.9): when the learning set is small, the computed model is flexible enough to reproduce almost all the learning set despite the errors; however, when the size of the learning set is large, as the computed model is more specific, the proportion of alternatives in the learning set whose assignment is not reproduced by the inferred model corresponds to the proportion of errors introduced in the learning set. Note however that alternatives in the learning set that are excluded when inferring the model do not necessarily correspond to the errors introduced in the learning set. However, the proportion of alternatives excluded when inferring the model is at most equal to the proportion of introduced errors. Also, it should be noted that the distribution of the restoration rate becomes more and more tightly grouped around its central value.

- The number of criteria $|\mathcal{N}|$: Figs. 13 (for U^B) and 14 (for U^C) show the variation of the restoration rate according to the

Fig. 12. Restoration rate by number of reference assignments (5 criteria, 3 categories and 10% noise) to learn a U^C model.

Fig. 13. Restoration rate by number of criteria (128 reference assignments, 3 categories and 10% noise) to learn a U^B model.

number of criteria. Increasing the number of criteria makes the problem more flexible, and consequently noticeably enhances the restoration rate with a convergence towards 100%

- The noise rate μ: Figs. 15 and 16 indicate that the restoration rate decreases linearly with the noise rate.

6.3.2. Computing time

Tables 8 and 9 show the distribution of the computation time in the baseline configuration (128 reference assignments, 5 crite-

ria, 3 categories and 10% noise) to learn both NCS models (U^B and U^C). When dealing with our baseline, applying the strategy based on the explicit representation of coalitions is 20 times faster than applying the strategy based on pairwise separation of alternatives, while this advantage was only threefold for the decision approaches (see e.g. Figs. 7 and 8): the relaxation from SAT to MaxSAT seems to favor the strategy based on coalitions.

We investigate the influence of the parameters describing the instance.
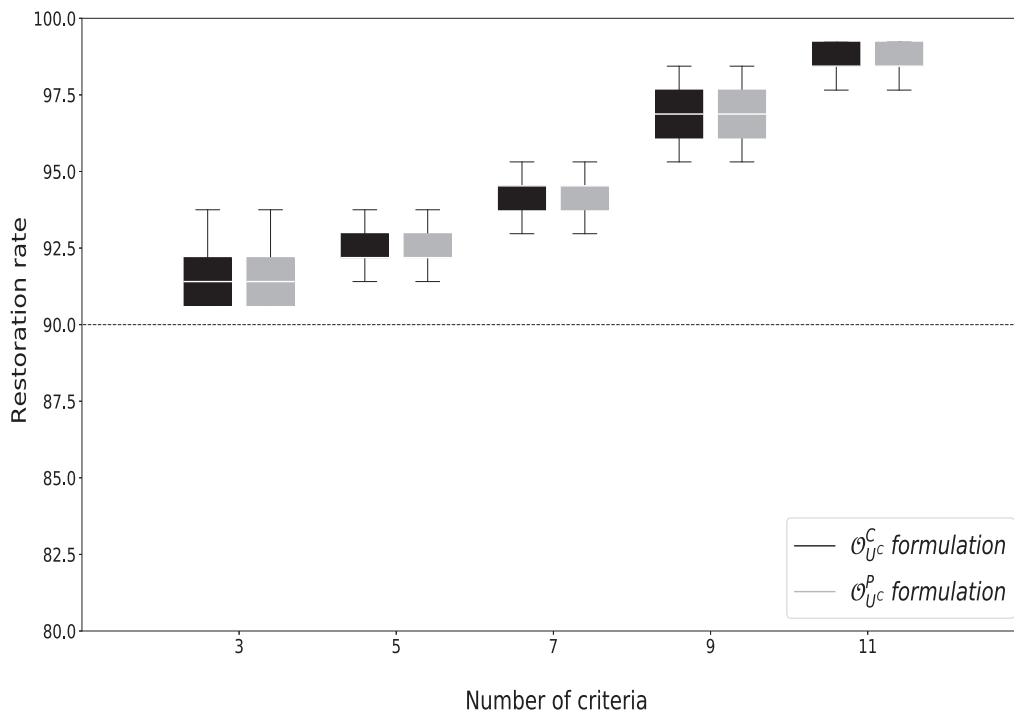
Fig. 14. Restoration rate by number of criteria (128 reference assignments, 3 categories and 10% noise) to learn a U^C model.
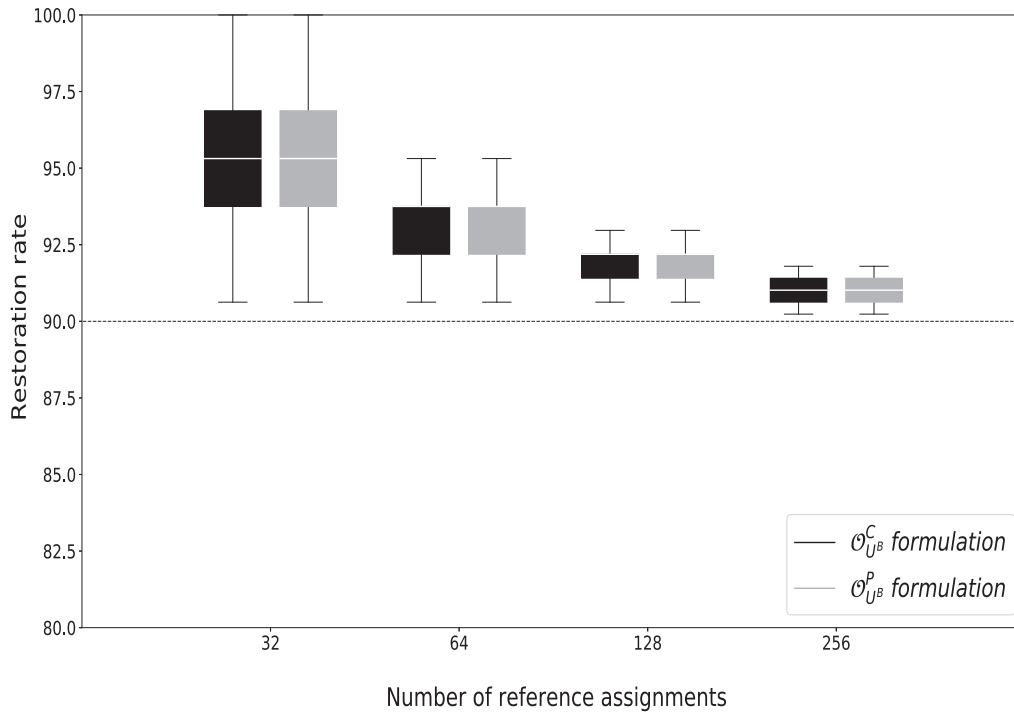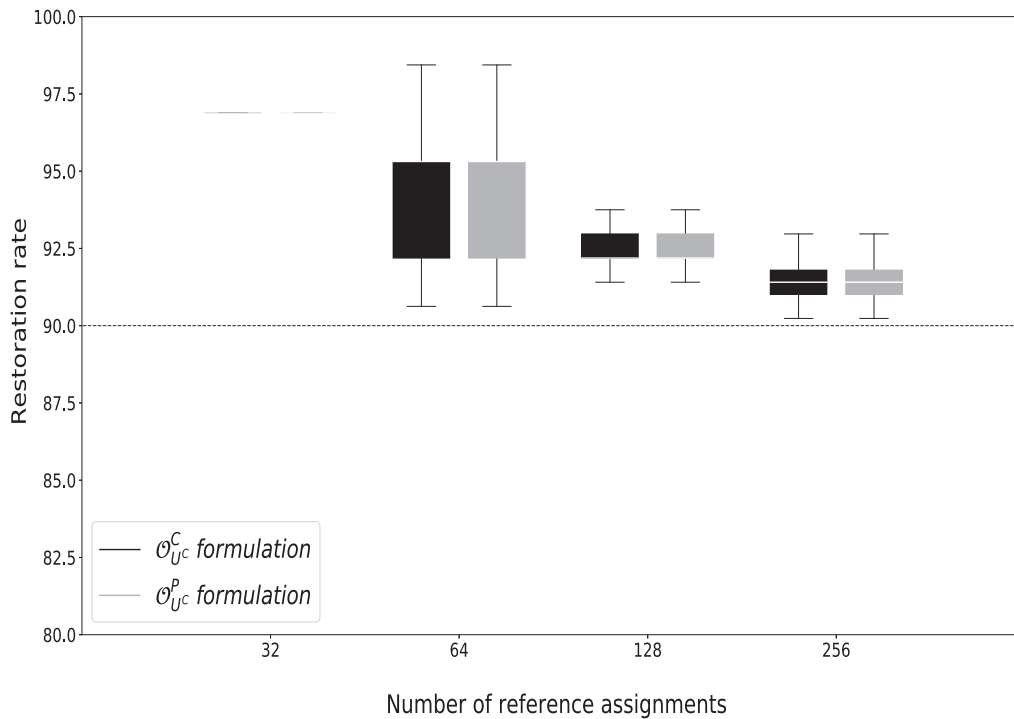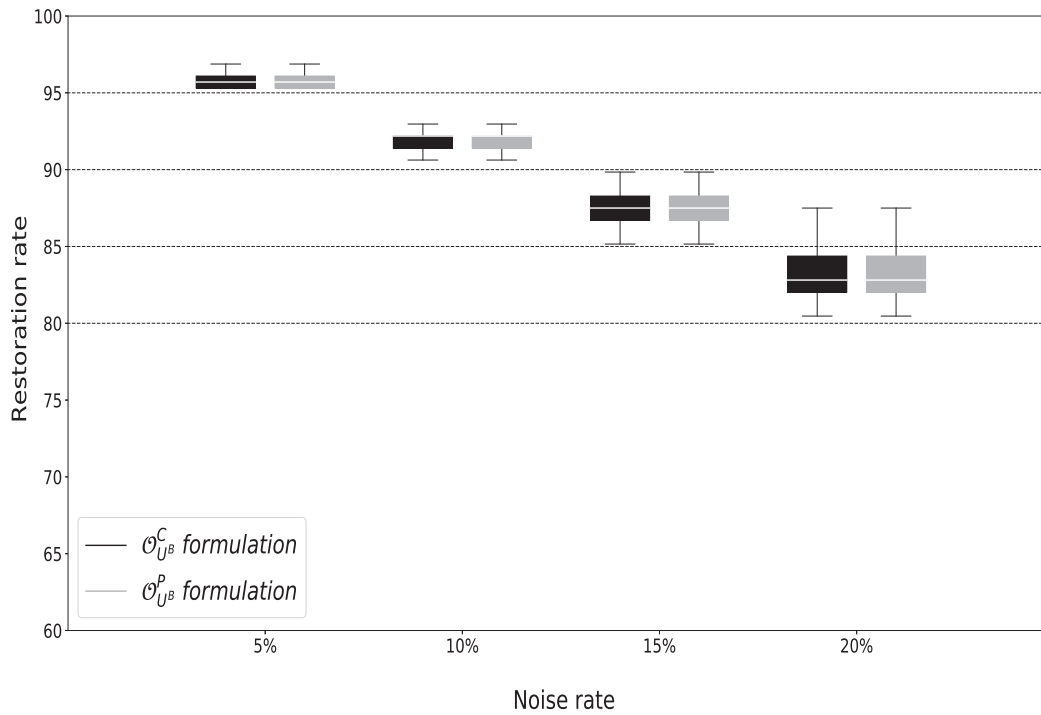

Fig. 15. Restoration rate by noise rate (128 reference assignments, 5 criteria and 3 categories) to learn a U^B model.

- The number of reference assignments $\mathbb{X}^*$: Figs. 17 and 18 indicate that the distribution of the computing time for the two MaxSAT-formulations and for both U^B and U^C models remains tightly grouped around its central value. It also shows that this value steadily increases with the number of reference assignments, consistently with the power laws found in Section 6.2.2, i.e. $t \propto |\mathbb{X}^*|$ for the representation strategy based on coalitions of criteria, and $t \propto |\mathbb{X}^*|^2$ for the

representation strategy based on the pairwise separation of alternatives.

- The number of criteria $|\mathcal{N}|$: Figs. 19 and 20 indicates that the distribution of the computing time, when applying both strategies and learning both models U^C and U^B, remains tightly grouped around its central value. These results remain consistent to the models proposed in Section 6.2.2: $t \propto |\mathcal{N}|$ for the representation strategy based on the pairwise separation of
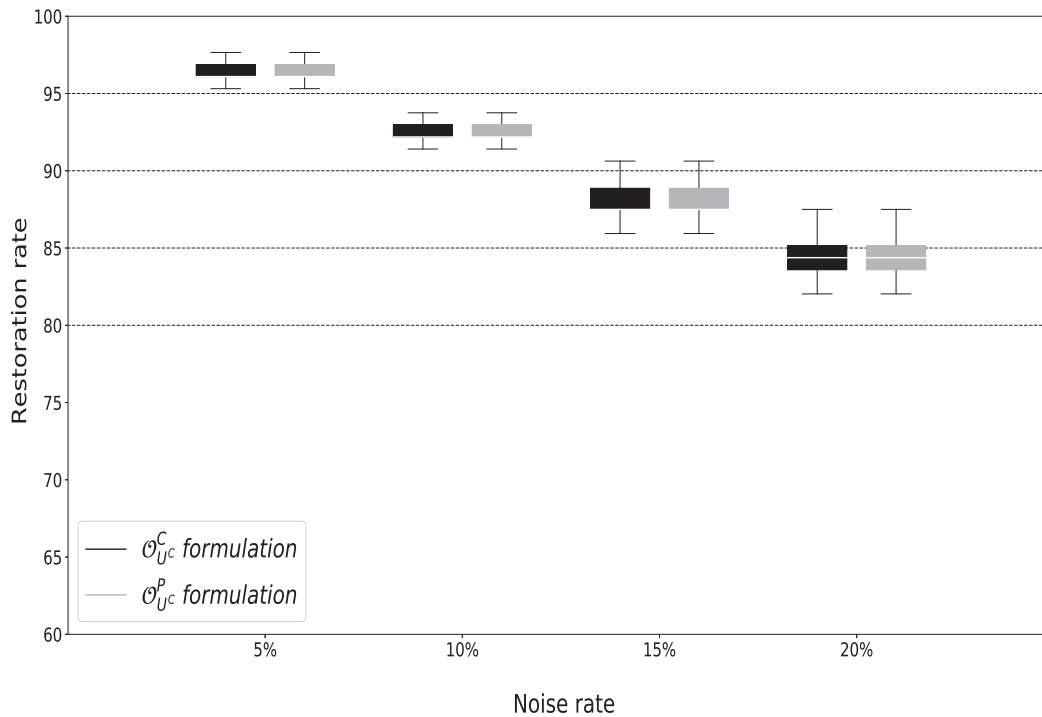
Fig. 16. Restoration rate by noise rate (128 reference assignments, 5 criteria and 3 categories) to learn a U^C model.
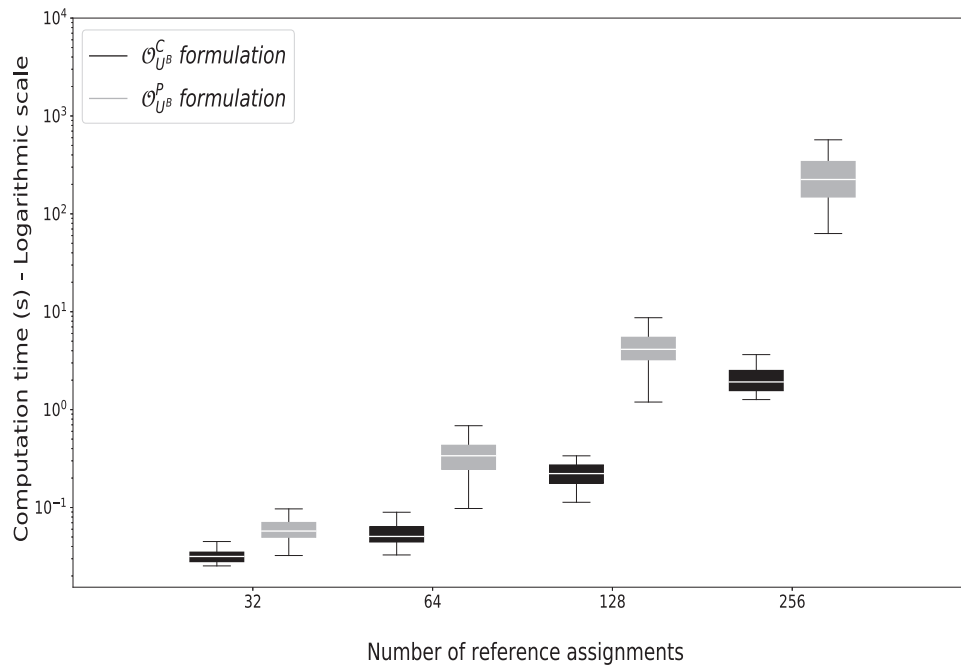
Fig. 17. Computation time by number of reference assignments (5 criteria, 3 categories, 10% noise) to learn a U^B model.

Table 8
Computation time to learn a U^B model in the baseline config. (128 ref. alt., 5 crit., 3 categ. and 10% noise).

	$\mathcal{O}^C_{U^B}$	$\mathcal{O}^P_{U^B}$
Max	0.337s	8.690s
2nd quartile	0.272s	5.492s
Median	0.222s	4.132s
1st quartile	0.176s	3.228s
Min	0.113s	1.195s

alternatives, $t \propto 2^{|\mathcal{N}|}$ for the representation strategy based on coalitions of criteria.

- The noise rate μ: The distribution of the computation time for both MaxSAT formulations remains tightly grouped around its central value, and $\log t$ increases linearly (with a low slope) with the noise rate (Figs. 21 and 22).

6.3.3. Results on the ability of the inferred model to restore the original one

The observations made in Section 7, concerning the irresoluteness of the approaches implementing the representation
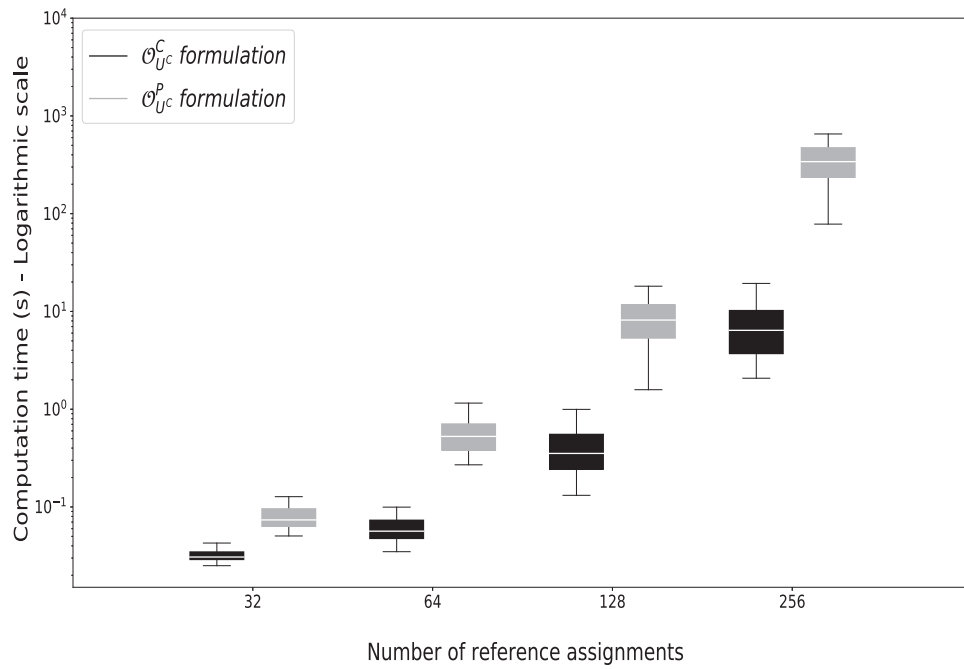
Fig. 18. Computation time by number of reference assignments (5 criteria, 3 categories, 10% noise) to learn a U^C model.
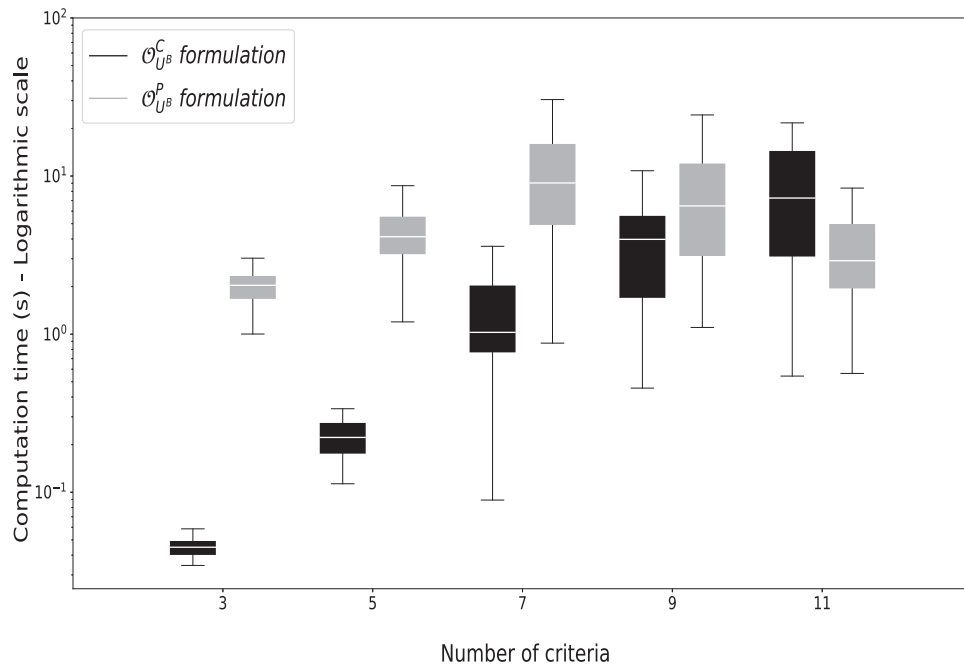
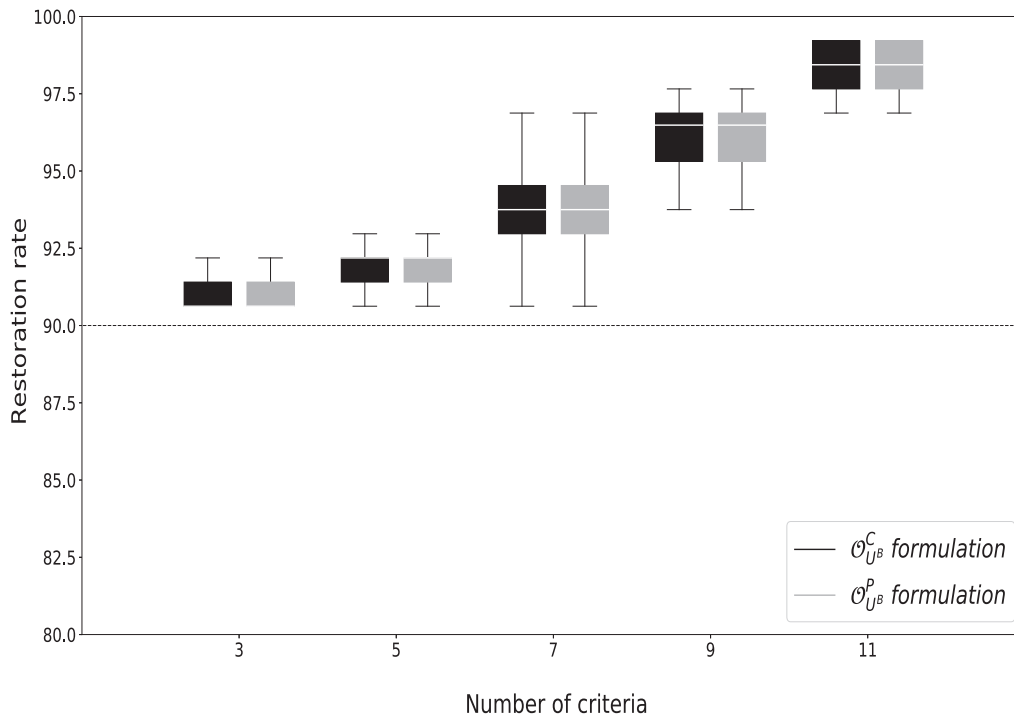**Fig. 19.** Computation time by number of criteria (128 reference assignments, 3 categories and 10% noise) to learn a U^B model.

strategy based on pairwise separation, remain valid when considering MaxSAT relaxations. Adopting the same notations as the SAT formulations, T-Student tests show that for both models U^B and U^C the generalization index when $\mathcal{T} = \mathcal{T}_{min}$ is always at least as good as the other two variants regardless of the number of criteria, alternatives, categories and the noise rate (see for instance the baseline configuration Table 10). The rule of thumb proposed in Section 7 remains valid when transposed to optimization approaches implemented via a MaxSAT solver – the post-processing strategy $\mathcal{T} = \mathcal{T}_{min}$ yields the best results, and is the only one represented on the subsequent figures.

The first two columns of Table 10 depicts the distribution of the generalization index for both MaxSAT formulations for the baseline situation (128 reference assignments, 5 criteria, 3 categories and 10% noise) for learning a U^B model (respectively a U^C model). For both models, the two distributions are almost the same with a slight difference on the median.

Figs. 23–26 present the variations of the alignment of the computed U^B models (respectively U^C models) yielded by both MaxSAT formulations with the ground truth. For both NCS variants, the experimental results display a tendency towards a degradation of this alignment as the number of criteria or the number of categories
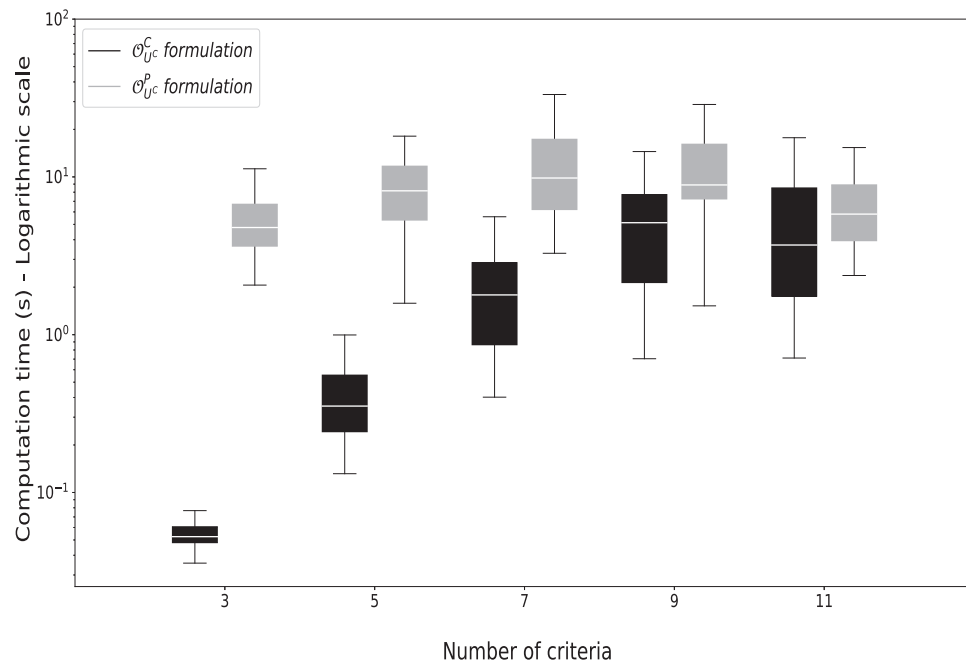
Fig. 20. Computation time by number of criteria (128 reference assignments, 3 categories and 10% noise) to learn a U^C model.

Fig. 21. Computation time by noise rate (128 reference assignments, 5 criteria and 3 categories) to learn a U^B model.

Table 9
Computation time to learn a U^C model in the baseline config.(128 ref. alt., 5 crit., 3 categ. and 10% noise).

	$\mathcal{O}^C_{U^C}$	$\mathcal{O}^P_{U^C}$
Max	0.996s	18.121s
2nd quartile	0.554s	11.7s
Median	0.352s	8.161s
1st quartile	0.242s	5.323s
Min	0.131s	1.582s

Table 10
Generalization index in the baseline configuration (128 reference assignments, 5 criteria, 3 categories and 10% noise) when learning a U^B model, for both representation strategies and three post-processing strategies.

	$\mathcal{O}^C_{U^B}$	$\mathcal{O}^P_{U^B}$		
		$\mathcal{T} = \mathcal{T}_{\min}$	$\mathcal{T} = \mathcal{T}_{rand}$	$\mathcal{T} = \mathcal{T}_{\max}$
Max	97.7%	99.2%	98.8%	98.8%
2nd quartile	95.6%	96.5%	95.9%	95.9%
Median	94.4%	95.4%	94.6%	94.5%
1st quartile	92.2%	93.5%	92.4%	92.6%
Min	87.7%	90%	88.2%	88%
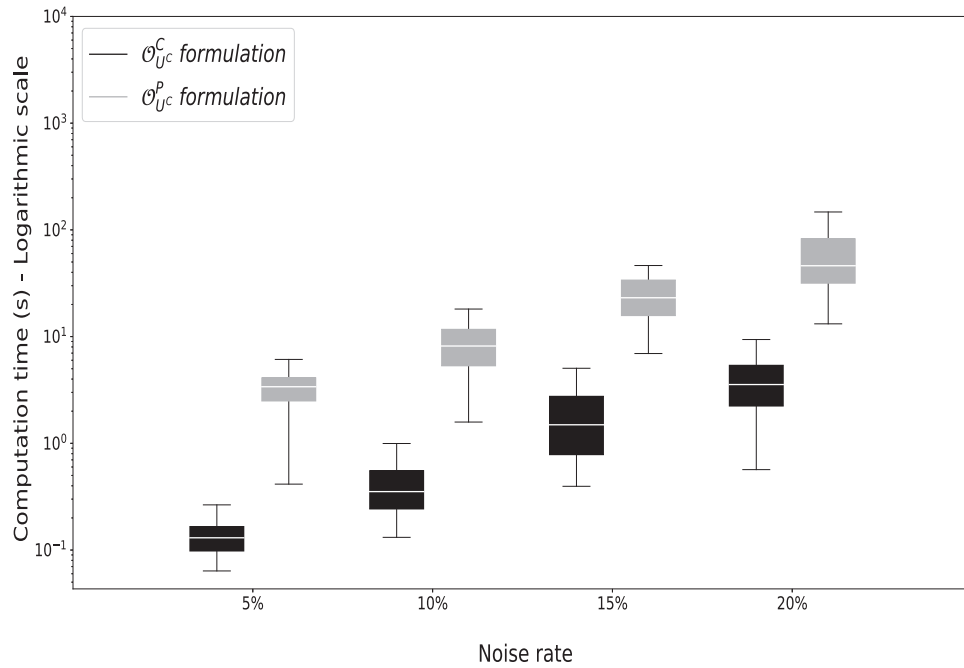
Fig. 22. Computation time by noise rate (128 reference assignments, 5 criteria and 3 categories) to learn a U^C model.
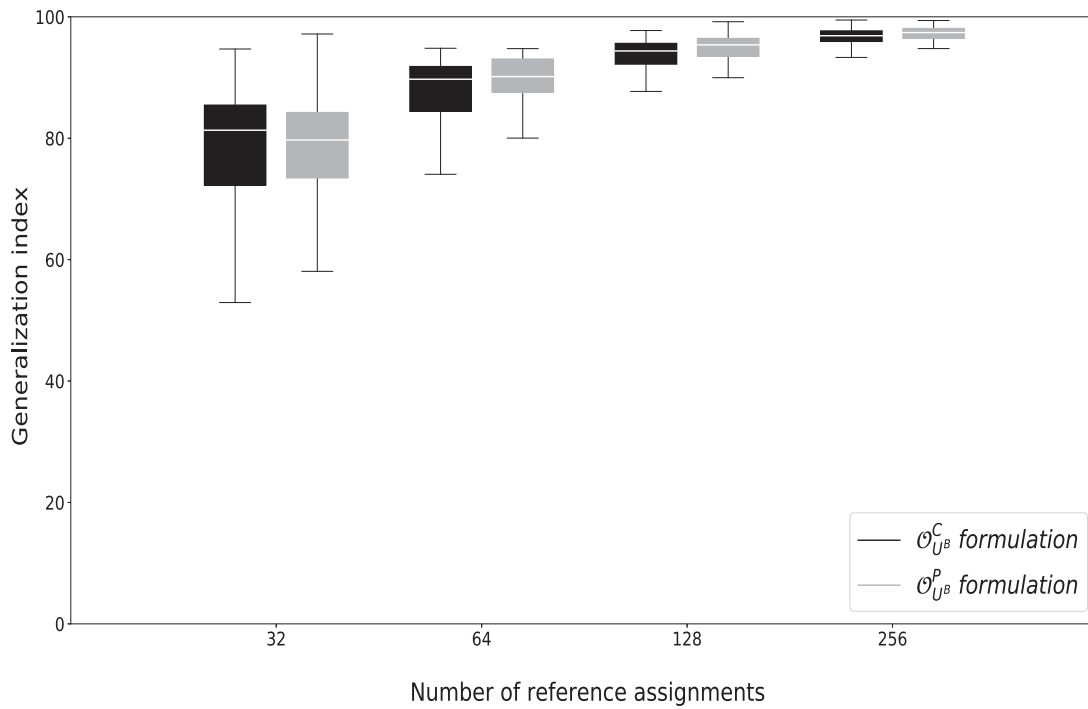
Fig. 23. Generalization index by number of reference assignments (5 criteria, 3 categories and 10% noise) to learn a U^B model.

increases. Conversely, as expected, increasing the number of reference assignments noticeably enhances the generalization index. The two formulations seem to behave in a similar manner with respect to the modification of these parameters. And finally, the generalization rate decreases linearly with the noise rate.

6.4. Discussion

In this section, we discuss the influence of input parameters (number of criteria, and the size of the learning set) on the computing time, the ability to restore learning sets, and to generalize of both representation strategies (the one based on explicit representation of coalitions, and the one based on pairwise separation of alternatives). The discussion focuses on both problem descriptions: decision (SAT) and optimization (MaxSAT) for learning both variants of NCS (U^B and U^C). The results obtained provides *(i)* the empirical confirmation of results which were expected, and *(ii)* insights for an analyst who wishes to use the proposed learning algorithms in an decision-aiding case study.

6.4.1. Empirical confirmation of expected results
Computation time:
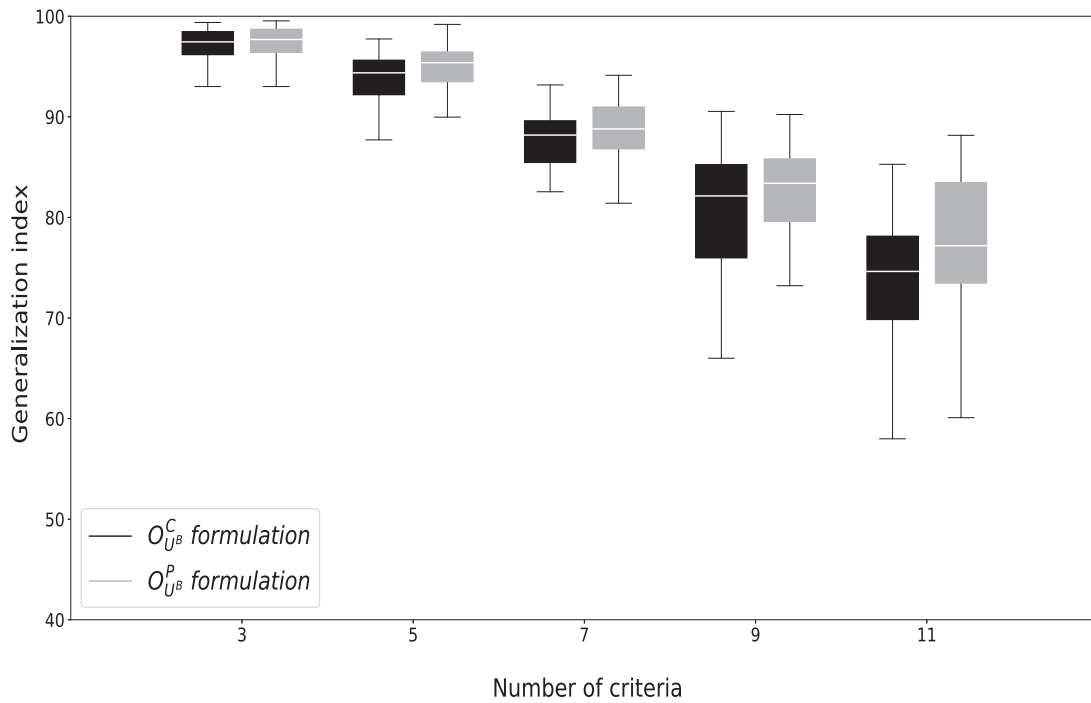
Fig. 24. Generalization index by number of criteria (128 reference assignments, 3 categories and 10% noise) to learn a U^B model.
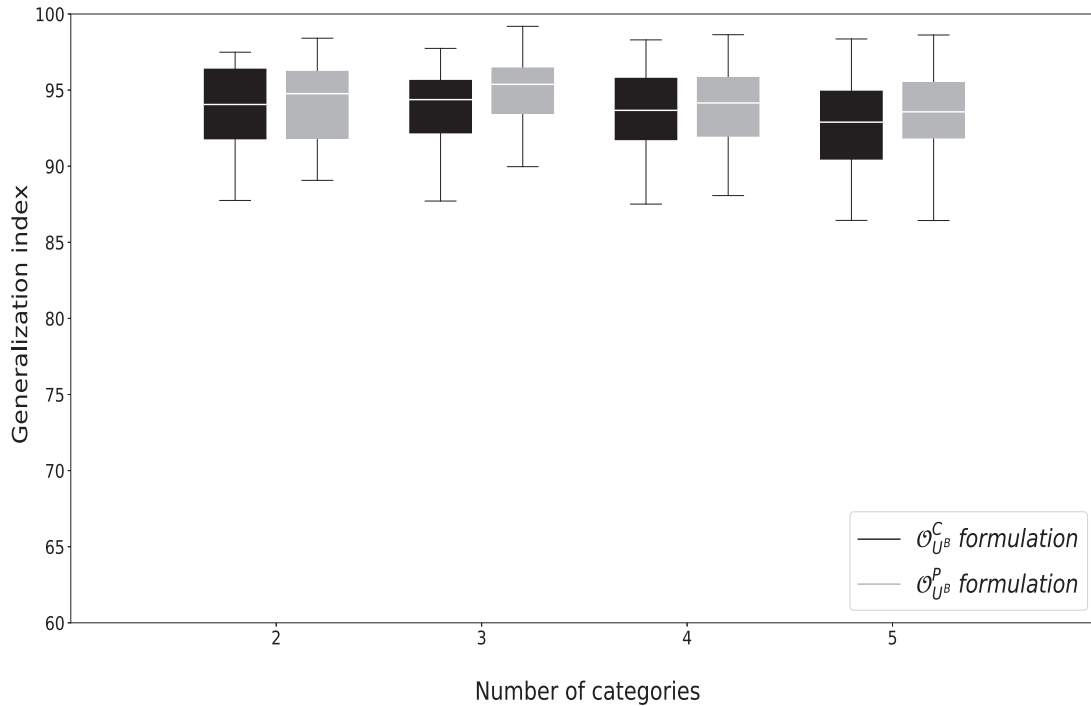
Fig. 25. Generalization index by number of categories (128 reference assignments, 5 criteria and 10% noise) to learn a U^B model.

On the one hand, for each NCS variants (U^B and U^C) and for both SAT and MaxSAT problem descriptions, the number of reference assignments impacts linearly the computation time of the coalitions-based representation strategy, and quadratically the computation time of the pairwise separation representation). On the other hand, the coalitions-based representation strategy depends exponentially on the number of criteria, and this dependence remains linear for the separation-based representation.

For a fixed number of criteria, when increasing the number of reference assignments, the coalition-based representation becomes faster than the separation-based representation (as the size of the learning set impacts the computing time linearly for the coalition-based representation, and quadratically for the separation-based representation).

Conversely, for a fixed number of reference assignments, when increasing the number criteria, the separation-based representation becomes faster than the coalition-based representation (as the number of criteria impacts the computing time exponentially for the coalition-based representation, and linearly for the separation-based representation).
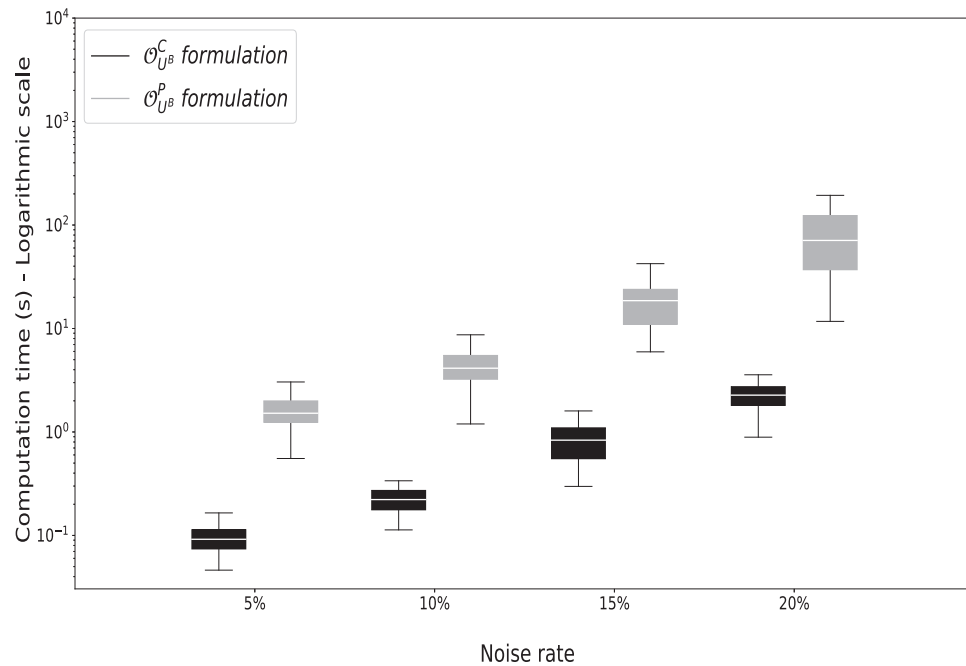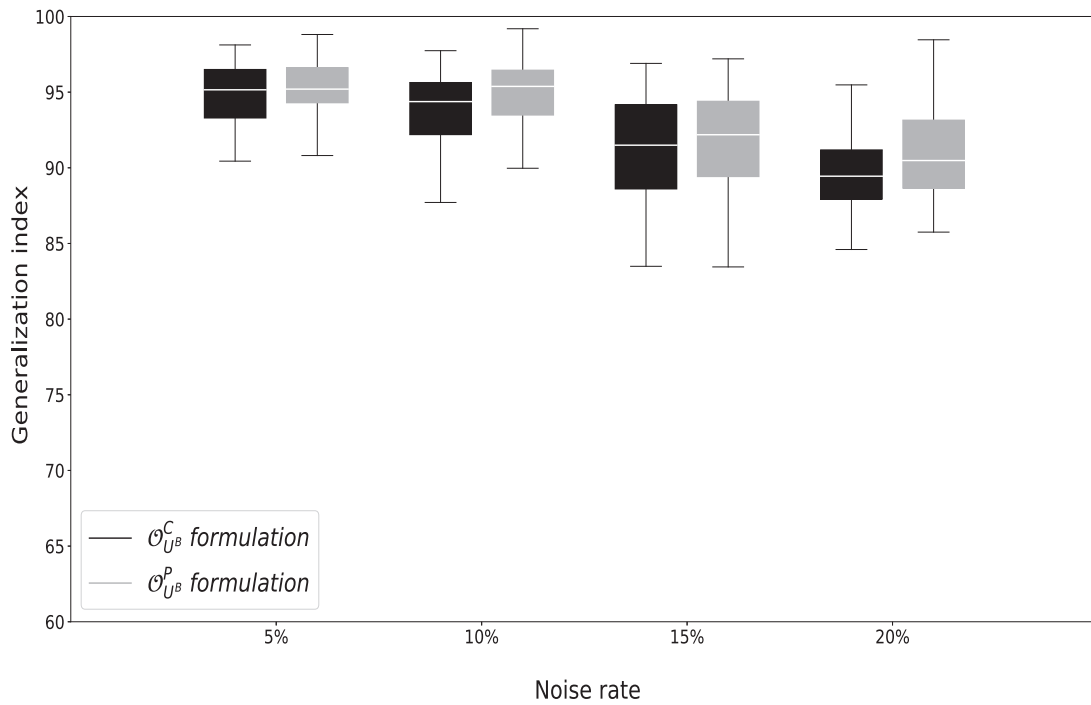
Fig. 26. Generalization index by noise rate (128 reference assignments, 5 criteria and 3 categories) to learn a U^B model.
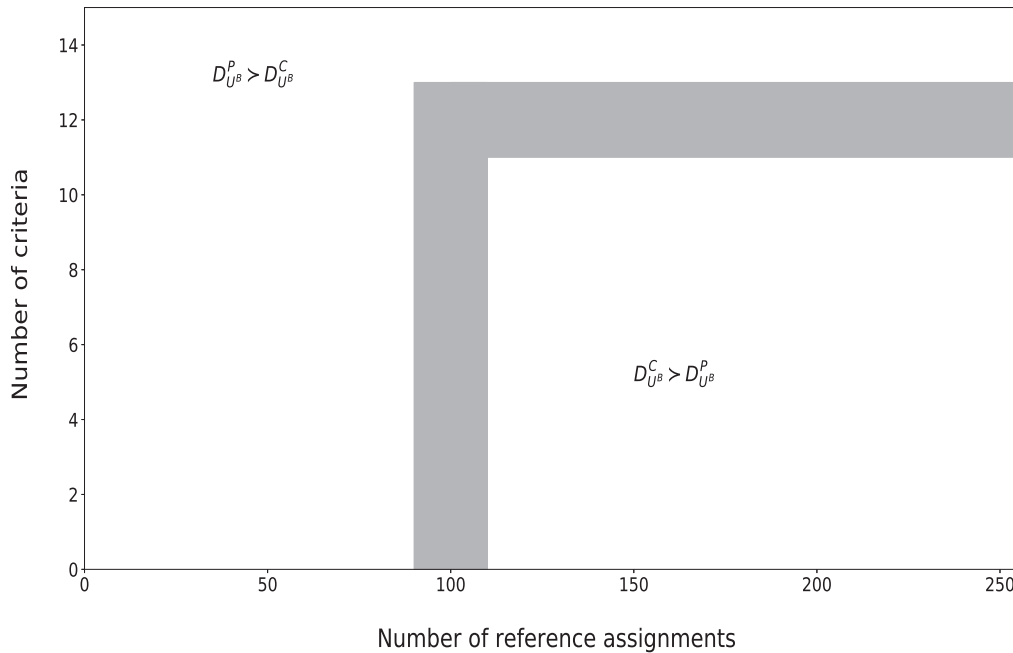
Fig. 27. Computation time of SAT problems by number of reference assignments and number of criteria (3 categories) to learn a U^B model.

These two effects leads to distinguish configurations (depending on the number of criteria, and size of the learning set) in which either of two representations (coalition-based or separation-based) is faster. For a decision problem with 3 categories (and 10% noise for MaxSAT instances), Figs. 27 and 28 (Figs. 29 and 30, respectively) depicts which of the two representations (coalition-based or separation-based) is faster to learn U^B model (U^C model, respectively) depending on the number of alternatives and number of criteria (both for SAT and MaxSAT). These Figures offer insights to choose the appropriate representation according to the number of criteria and the number of alternatives.

Ability to restore the learning set:

As expected, all SAT instances (without noise) are able to fully restore the learning sets; this result is an experimental validation of the theoretical work developed in Section 4. Moreover, when learning a model from noisy learning sets (MaxSAT extension), we were able to infer NCS models with a restoration rate over $1 - x$, where x denotes the noise level in the learning set.

Ability to generalize:

In terms of generalization (the alignment between the output model with the ground truth), for both U^B and U^C models,
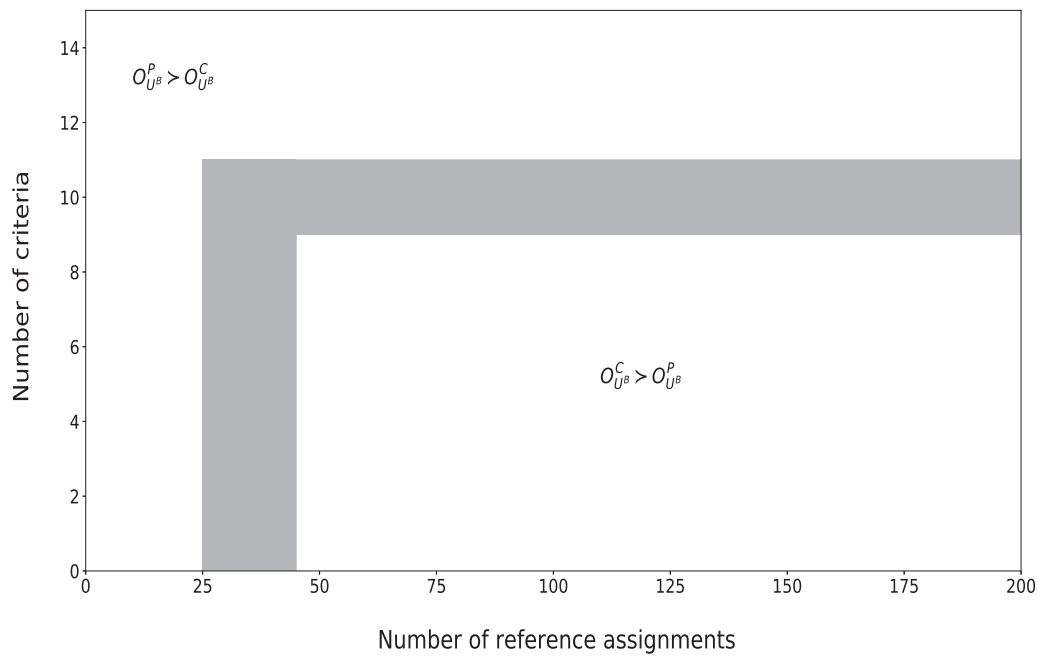
Fig. 28. Computation time of MaxSAT problems by number of reference assignments and number of criteria (3 categories and 10% noise) to learn a U^B model.
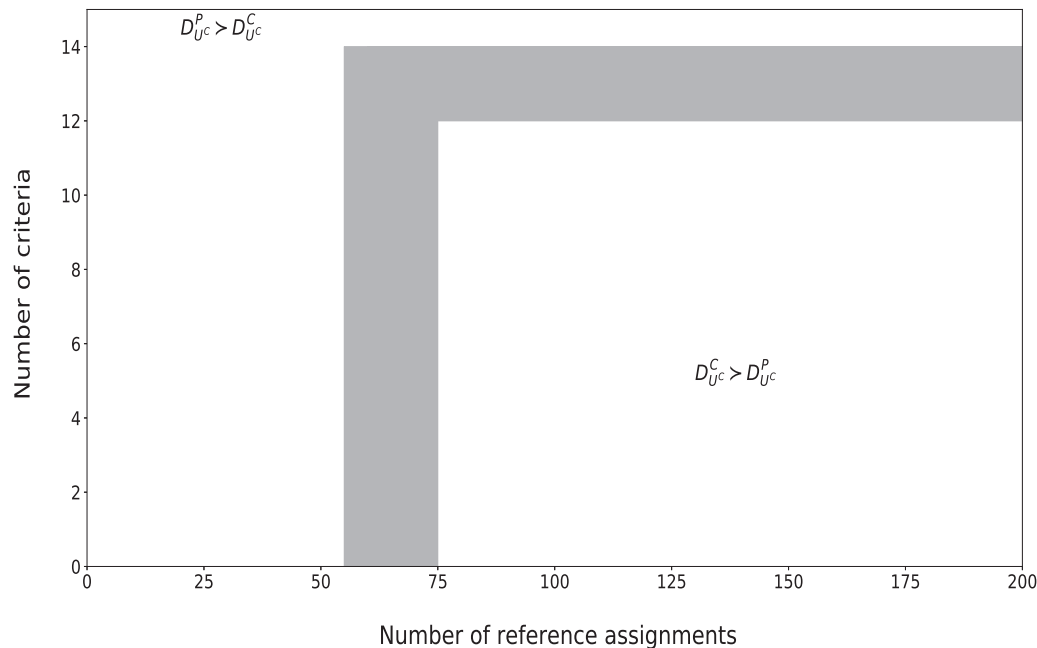
Fig. 29. Computation time of SAT problems by number of reference assignments and number of criteria (3 categories) to learn a U^C model.

coalition-based and separation-based strategies behave in analogously:

- an increase of the size of the learning set induces an improvement of the generalization index; such improvement occurs whatever the noise level (up to 20%). This means that it seems always possible to "capture the ground truth" with a sufficiently large learning set,
- an increase in the reference set noise level requires a larger learning set to keep the same generalization level. This implies that the "quality" of the learning set, have a significant impact on the required size of this learning set.

6.4.2. Insights for the decision analyst

An interesting aspect of the empirical results lies in the possibility to derive insights on how to put the proposed learning algorithms in practice in an decision-aiding case study.

Defining the size of the learning set for a given number of criteria:

An important question for a decision analyst concerns the number of assignment examples to collect in order to accurately capture the DM's preferences. Our experiments provide figures to answer such questions. In a decision problem involving 3 categories and 5 criteria, if the analyst wishes to obtain an U^B model with
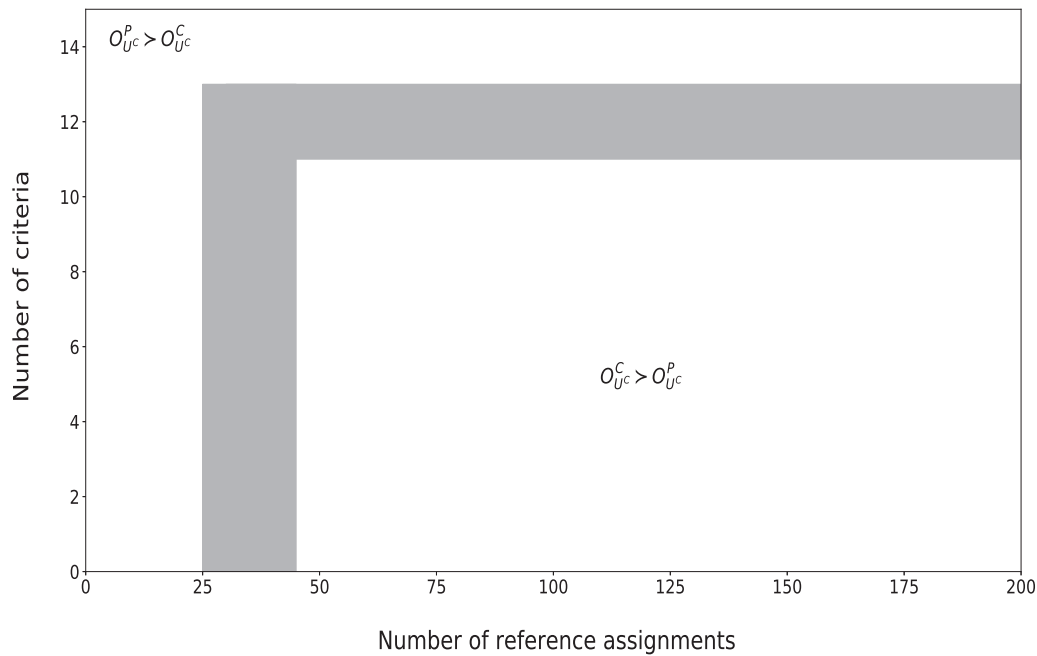
Fig. 30. Computation time of MaxSAT problems by number of reference assignments and number of criteria (3 categories and 10% noise) to learn a U^C model.

target level of 90% for the generalization index, and postulates an error rate of 10% in the set of assignment examples, Fig. 23 informs us that the size of the learning set should be in the interval [64, 128].

Choosing the fastest formulation depending on the number of criteria and size of the learning set:

Another relevant question concerns which of the coalition-based or separation-based representation provides the lowest computing time for a given size of learning set (and number of criteria).

For a given number of criteria and for learning a U^B model, Figs. 27 and 28 depict the approximate thresholds in terms of number of reference assignments from which the coalition-based representation becomes faster than the separation-based one. In the case where the preference information is perfect and for less than ∼50 examples, the separation-based representation is faster than the coalition-based representation, and the generalization is equivalent for both representations. For MaxSAT instances, for more than ∼50 examples and less than ∼11 criteria, the coalition-based representation formulation is faster than the separation-based one. However, for a number of criteria exceeding ∼13 or for less than ∼50 reference assignments, the separation-based representation is faster. For all configurations, the separation-based representation generalizes better.

For a given number of criteria and for learning a U^C model, Figs. 29 and 30 depict the approximate thresholds in terms of number of reference assignments from which the coalition-based representation becomes faster than the separation-based one. In the case where the preference information is noiseless and for more than ∼14 criteria or for less than ∼64 reference assignments, the separation-based representation is more efficient than the coalition-based one in terms of the computation time and the generalization index.

7. Conclusion

In this paper, we consider the multiple criteria Non-Compensatory Sorting model and its variants with a unique profile

(U^B) and a unique set of sufficient coalitions (U^C). Learning this model has already been addressed by the literature, and solved by the resolution of a MIP (Leroy et al., 2011) or via a specific heuristic (Sobrie, Mousseau, & Pirlot, 2013; 2015). Recently, two SAT representations (coalition-based, and separation-based) have been proposed to learn such a model from perfect preference information and already proved to be superior to other approaches, see Belahcene et al. (2018b). The separation-based representation was originally described in Belahcene et al. (2018b) but only focusing on the case with two categories. We consider in this work the generalization of this formulation to the multiple categories case for learning NCS and its variant U^B-NCS and U^C-NCS. The separation-based representation is more compact than the coalition-based one as it handles explicitly a set of sufficient coalitions that lies in the power set of the criteria. In order to handle the inconsistency in preference information, we extend the two SAT problems using MaxSAT language. Thus, for each variant of NCS, we proposed two MaxSAT programs to compute the model's parameters from noisy preference information.

The separation-based representation proposed for learning U^B and U^C models is at least as good as the coalition-based one in terms of generalization and for both types of preference information (perfect and not-so-perfect preferences). Computation time of the two representations evolves depending on the number of reference alternatives and the number of criteria; the separation-based representation performs better when the number of criteria increases, while it is not the case when the number of reference alternatives increases. Increasing the number of categories penalizes the separation-based representation proposed for learning U^B model, since the number of clauses depends quadratically on the number of categories.

However, for real world decision problems, assuming that the number of reference assignments is ∼100 examples, we can consider two types of applications: an application that involves a large number of criteria ($|\mathcal{N}| > \sim 12$) and therefore the separation-based representation seems better as it is faster and generalizes better than the first one, and an application that involves a limited number of criteria ($|\mathcal{N}| < \sim 10$), in this case, the coalition-based repre-

sentation is slightly faster and generalizes less than the separation-based one.

Finally, our work shows that, when learning MCDA models from preference information, SAT and MaxSAT languages can be relevant and efficient. This is specifically the case for ordinal MCDA aggregation procedures based on pairwise comparison of alternatives (so called outranking methods, see Figueira, Mousseau, & Roy (2005)). We believe that our work opens avenue for further research to develop new algorithms to learn outranking models from preference statements using SAT/MaxSAT language.

Appendix A. SAT formulations for NCS variants with more than 2 categories

A1. Learning a U^B-NCS model

When trying to fit a U^B-NCS model, neither a variables nor s variables are indexed by exigence level; s variables are indexed by a criterion i and a pair of alternatives $g, b \in \mathbb{X}^\star$ such that g is preferred to b, i.e. $\alpha(g) \succ \alpha(b)$.

The propositional formula obtained by following the representation strategy based on the pairwise separation of alternatives is particularly simple and elegant.

Definition A.1. Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$, we define the boolean function $\Phi_\alpha^{P'B}$ with variables $\langle a_{i,x} \rangle_{i \in \mathcal{N},\ k \in [2.p], x \in \mathbb{X}^\star}$ and $\langle s_{i,g,b} \rangle_{i \in \mathcal{N},\ \alpha(g) \succ \alpha(b)}$, as the conjunction of clauses:

$$\Phi_{\alpha, U^B-NCS}^{P'B} = \phi_\alpha^{P'B1} \wedge \phi_\alpha^{P'B3} \wedge \phi_\alpha^{P'B4} \wedge \phi_\alpha^{P'B5}$$

$$\phi_\alpha^{P'B1} = \bigwedge_{i \in \mathcal{N}} \bigwedge_{x' \succsim_i x \in \mathbb{X}^\star} (a_{i,x'} \vee \neg a_{i,x})$$
$$\phi_\alpha^{P'B3} = \bigwedge_{i \in \mathcal{N}} \bigwedge_{\alpha(g) \succ \alpha(b)} (\neg s_{i,g,b} \vee \neg a_{i,b})$$
$$\phi_\alpha^{P'B4} = \bigwedge_{i \in \mathcal{N}} \bigwedge_{\alpha(g) \succ \alpha(b)} (\neg s_{i,g,b} \vee a_{i,g})$$
$$\phi_\alpha^{P'B5} = \bigwedge_{\alpha(g) \succ \alpha(b)} (\bigvee_{i \in \mathcal{N}} s_{i,g,b})$$

Corollary A.1. *Given a contex, an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$ can be represented in the Non-Compensatory sorting model with unique profile if, and only if, $\Phi_\alpha^{P'B}$ is satisfiable.*

This condition is obviously necessary. It is sufficient because the sets of observably sufficient and insufficient coalitions are nested by construction, even in the case $\mathcal{A}_i^2 = \cdots = \mathcal{A}_i^p$.

A2. Learning a U^C-NCS model

We describe here the generalization of the pairwise separation formulation Φ_α^P (see Definition 4.3) to the multiple category case for fitting a U^C-NCS (Unique set of sufficient coalitions) model. Given a nesting of approved sets $\langle \mathcal{A}_i^h \rangle$, this unique set of sufficient coalitions satisfies all the constraints put by the observed sufficient and insufficient coalitions of criteria at every exigence level. This observation yields the following lower and upper bounds:

$$\mathcal{S}_{\langle \mathcal{A}_i^k \rangle}(\alpha) = Cl_{\mathcal{P}(\mathcal{N})}^{\supseteq}(\bigcup_{k \in [2.p]} \bigcup_{g \in \alpha^{-1}(C^{\geq k})} \{i \in \mathcal{N} : g \in \mathcal{A}_i^k\})$$

$$\mathcal{F}_{\langle \mathcal{A}_i^k \rangle}(\alpha) = Cl_{\mathcal{P}(\mathcal{N})}^{\subseteq}(\bigcup_{k \in [2.p]} \bigcup_{b \in \alpha^{-1}(C^{\prec k})} \{i \in \mathcal{N} : b \in \mathcal{A}_i^k\})$$

In turn, this entails a modification of the third condition (pairwise separation) of the representation theorem (Theorem 4.2):

3C. (*pairwise separation for a unique set of sufficient coalitions*) for each exigence levels $k \in [2.p]$ and $k' \in [2.p]$, for each pair of alternatives $(g, b) \in (\mathbb{X}^*)^2$ such that $g \in \alpha^{-1}(C^{\geq k'})$ and $b \in \alpha^{-1}(C^{\prec k})$, there is at least one point of view $i \in \mathcal{N}$ such that $g \in \mathcal{A}_i^{k'}$ and $b \notin \mathcal{A}_i^k$.

We translate this modified representation theorem into a SAT formulation equisatisfiable with Inv-U^B-NCS, using variables a indexed by a criterion, an exigence level and a reference alternative, as well as variables s indexed by a criterion, a pair of exigence levels, and a pair of alternatives.

Definition A.2. Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$, we define the boolean function $\Phi_\alpha^{P'C}$ with variables $\langle a_{i,k,x} \rangle_{i \in \mathcal{N},\ k \in \{2.p\}, x \in \mathbb{X}^\star}$ and

$\langle s_{i,k,k',g,b} \rangle_{i \in \mathcal{N},\ k \in \{2.p\},\ k' \in \{2.p\},\ g \in C^{\geq k'},\ b \notin C^{\geq k}}$, as the conjunction of clauses:

$$\Phi_\alpha^{P'C} = \phi^{P'1} \wedge \phi^{P'2} \wedge \phi^{P'C3} \wedge \phi^{P'C4} \wedge \phi^{P'C5}$$

$$\phi_\alpha^{P'C3} = \bigwedge_{i \in \mathcal{N},\ k \in [2.p],\ k' \in [2.p]} \bigwedge_{g \in \alpha^{-1}(C^{\geq k'}),\ b \in \alpha^{-1}(C^{\prec k})} (\neg s_{i,k,k',g,b} \vee \neg a_{i,k,b})$$
$$\phi_\alpha^{P'C4} = \bigwedge_{i \in \mathcal{N},\ k \in [2.p],\ k' \in [2.p]} \bigwedge_{g \in \alpha^{-1}(C^{\geq k'}), b \in \alpha^{-1}(C^{\prec k})} (\neg s_{i,k,k',g,b} \vee a_{i,k',g})$$
$$\phi_\alpha^{P'C5} = \bigwedge_{k \in [2.p],\ k' \in [2.p]} \bigwedge_{g \in \alpha^{-1}(C^{\geq k'}),\ b \in \alpha^{-1}(C^{\prec k})} (\bigvee_{i \in \mathcal{N}} s_{i,k,k',g,b})$$

Formulations of $\phi_\alpha^{P'1}$ and $\phi_\alpha^{P'2}$ can be found in Definition 4.4.

Corollary A.2. *Given a context, an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$ can be represented in the Non-Compensatory sorting model with a unique set of sufficient coalitions if, and only if, $\Phi_\alpha^{P'C}$ is satisfiable.*

Appendix B. MaxSAT relaxations based on pairwise separation conditions for more than two categories

We provide here extensions of the MaxSAT formulation presented in Section 5.2, to the case with multiple categories. They rely on the fact that an *NCS* model with p categories is informally the combination of $p - 1$ *NCS* models with two categories whose parameters satisfy the nesting conditions on coalitions and satisfactory values. The maximization of the restoration in the second formulation is equivalent to the simultaneous maximization of the restoration in the sub-problems with two categories. On top of the 'z' variables encoding the correct restoration of a reference alternative, we introduce intermediate switches:

- 'y' variables, indexed by an alternative $x \in \mathbb{X}^\star$ and an exigence level $k \in [2.p]$, encode the proper restoration of alternative x by the 2-categories NCS model with GOOD$= C^{\geq k}$ and BAD$= C^{\prec k}$.

These variables are logically tied to the 'z' variables by the following conjunction of hard clauses:

$$\phi_\alpha^{\widetilde{P'yz}} = \bigwedge_{x \in \mathbb{X}^\star} \bigwedge_{k \in [2.p]} (y_{k,x} \vee \neg z_x)$$

While the objective in the MaxSAT formulation is to maximize the number of properly classified alternatives, this goal is reached by the simultaneous maximization of the restoration rate in each sub-problem with two categories, leading to the introduction of a number of sub-goals:

$$\phi_\alpha^{subgoals} = \bigwedge_{k \in [2.p]} \bigwedge_{x \in \mathbb{X}^\star} y_{k,x}$$

The soft clause ϕ_α^{goal} is given weight w_1, and each one of the clause appearing in the conjunction $\phi_\alpha^{subgoals}$ is given weight w_2, while the hard clauses are given weight $w_{\max}$. These weights are chosen so that $w_{\max} \gg w_1 \gg w_2$, and more precisely : $(p-1)|\mathbb{X}^\star|w_2 < w_1$; and $|\mathbb{X}^\star|w_1 < w_{\max}$.

The hard clauses differ according to the target model.

B1. Learning an NCS model

Use the following conjunction of hard clauses: $\phi_\alpha^{P'1} \wedge \phi_\alpha^{P'2} \wedge \phi_\alpha^{\widetilde{P'3}} \wedge \phi_\alpha^{\widetilde{P'4}} \wedge \phi_\alpha^{\widetilde{P'5}} \wedge \phi_\alpha^{\widetilde{P'yz}}$.

$$\phi_\alpha^{\widetilde{P'3}} = \bigwedge_{i\in\mathcal{N},\ 2\leq k\leq k'\leq p} \bigwedge_{g\in\alpha^{-1}(C^{\geq k'}),\ b\in\alpha^{-1}(C^{\lessdot k})} (\neg s_{i,k,k',g,b} \vee \neg a_{i,k,b})$$

$$\phi_\alpha^{\widetilde{P'4}} = \bigwedge_{i\in\mathcal{N},\ 2\leq k\leq k'\leq p} \bigwedge_{g\in\alpha^{-1}(C^{\geq k'}),b\in\alpha^{-1}(C^{\lessdot k})} (\neg s_{i,k,k',g,b} \vee a_{i,k',g})$$

$$\phi_\alpha^{\widetilde{P'5}} = \bigwedge_{k\in[2.p],2\leq k\leq k'\leq p} \bigwedge_{g\in\alpha^{-1}(C^{\geq k'}),\ b\in\alpha^{-1}(C^{\lessdot k})}$$
$$\left(\bigvee_{i\in\mathcal{N}} s_{i,k,k',g,b} \vee \neg y_{k,b} \vee \neg y_{k',g}\right)$$

Note that the conjunction $\phi_\alpha^{\widetilde{P'3}}$ (resp. $\phi_\alpha^{\widetilde{P'4}}$) subsumes the conjunction $\phi_\alpha^{P'3}$ (resp. $\phi_\alpha^{P'4}$) introduced in Definition 4.4, but that, together with the constraints $\phi_\alpha^{P'2}$, is equivalent to it. While this redundancy is not needed in the SAT formulation, it helps formulate the subgoals of the MaxSAT formulation.

B2. Learning a U^C-NCS model

Use the following conjunction of hard clauses: $\phi_\alpha^{P'1} \wedge \phi_\alpha^{P'2} \wedge \phi_\alpha^{P'C3} \wedge \phi_\alpha^{\widetilde{P'C4}} \wedge \phi_\alpha^{\widetilde{P'C5}} \wedge \phi_\alpha^{\widetilde{P'yz}}$.

Formulas $\phi_\alpha^{P'1}$ and $\phi_\alpha^{P'2}$ are introduced in Definition 4.4, $\phi_\alpha^{P'C3}$ and $\phi_\alpha^{P'C4}$ are introduced in Definition A.2, and

$$\phi_\alpha^{\widetilde{P'C5}} = \bigwedge_{k\in[2.p],\ k'\in[2.p]} \bigwedge_{g\in\alpha^{-1}(C^{\geq k'}),\ b\in\alpha^{-1}(C^{\lessdot k})}$$
$$\times \left(\bigvee_{i\in\mathcal{N}} s_{i,k,k',g,b} \vee \neg y_{k,b} \vee \neg y_{k',g}\right).$$

B3. Learning a U^B-NCS model

As it is the case when addressing the decision problem, the U^B-NCS model can be learned with a MaxSAT formulation which is very close to the one used in the case of two categories, without using any 'y' variables. Use the following conjunction of hard clauses (each one with weight $w_{\max}$): $\phi_\alpha^{P'B1} \wedge \phi_\alpha^{P'B3} \wedge \phi_\alpha^{P'B4} \wedge \phi_\alpha^{\widetilde{P'B5}}$, together with the soft clause ϕ_α^{goal} with weight $w_1 < w_{\max}/|\mathbb{X}^\star|$. Formulas $\phi_\alpha^{P'B1}$, $\phi_\alpha^{P'B3}$ and $\phi_\alpha^{P'B4}$ can be found in Definition. A.1, and

$$\phi_\alpha^{\widetilde{P'B5}} = \bigwedge_{\alpha(g)\succ\alpha(b)} \left(\bigvee_{i\in\mathcal{N}} s_{i,g,b} \neg z_b \vee \neg z_g\right).$$

References

Almeida-Dias, J., Figueira, J. R., & Roy, B. (2010). Electre Tri-C: A multiple criteria sorting method based on characteristic reference actions. *European Journal of Operational Research, 204*(3), 565–580.

Almeida-Dias, J., Figueira, J. R., & Roy, B. (2012). A multiple criteria sorting method where each category is characterized by several reference actions: The Electre Tri-nC method. *European Journal of Operational Research, 217*(3), 567–579.

Arcidiacono, S. G., Corrente, S., & Greco, S. (2021). Robust stochastic sorting with interacting criteria hierarchically structured. *European Journal of Operational Research, 292*(2), 735–754.

Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., & Ouerdane, W. (2017a). A model for accountable ordinal sorting. In *Proceedings of the 26th international joint conference on artificial intelligence (IJCAI 2017)* (pp. 814–820).

Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., & Ouerdane, W. (2018a). An efficient SAT formulation for learning multiple criteria non-compensatory sorting rules from examples. *Computers & Operations Research, 97*, 58–71.

Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., Ouerdane, W., & Chevaleyre, Y. (2018b). Accountable approval sorting. In *Proceedings of the 27th international joint conference on artificial intelligence (IJCAI 2018)*.

Belahcene, K., Mousseau, V., Pirlot, M., & Sobrie, O. (2017b). Preference elicitation and learning in a multiple criteria decision aid perspective. *Technical Report*. Laboratoire Génie Industriel, CentraleSupélec. Research report 2017-02

Benabbou, N., Perny, P., & Viappiani, P. (2016). A regret-based preference elicitation approach for sorting with multicriteria reference profiles. In *Proceeding of the 2nd workshop from multiple criteria Decision aid to Preference Learning (DA2PL)*.

Benabbou, N., Perny, P., & Viappiani, P. (2017). Incremental elicitation of Choquet capacities for multicriteria choice, ranking and sorting problems. *Artificial Intelligence, 246*, 152–180. https://doi.org/10.1016/j.artint.2017.02.001.

Błaszczyński, J., Greco, S., & Słowiński, R. (2007). Multi-criteria classification – A new scheme for application of dominance-based decision rules. *European Journal of Operational Research, 181*(3), 1030–1044.

Bouyssou, D., & Marchant, T. (2007a). An axiomatic approach to noncompensatory sorting methods in MCDM, I: the case of two categories. *European Journal of Operational Research, 178*(1), 217–245.

Bouyssou, D., & Marchant, T. (2007b). An axiomatic approach to noncompensatory sorting methods in MCDM, II: more than two categories. *European Journal of Operational Research, 178*(1), 246–276.

Bouyssou, D., Marchant, T., & Pirlot, M. (2020). A theoretical look at ELECTRE TRI-nB. In *Working paper*. https://hal.archives-ouvertes.fr/hal-02917994

Chateauneuf, A., & Jaffray, J. Y. (1987). Derivation of some results on monotone capacities by mobius inversion. In B. Bouchon, & R. R. Yager (Eds.), *Uncertainty in knowledge-based systems* (pp. 95–102). Springer.

Corrente, S., Doumpos, M., Greco, S., Słowiński, R., & Zopounidis, C. (2017). Multiple criteria hierarchy process for sorting problems based on ordinal regression with additive value functions. *Annals of Operations Research, 251*(1), 117–139.

Devaud, J. M., Groussaud, G., & Jacquet-Lagreze, E. (1980). UTADIS: Une methode de construction de fonctions d'utilite additives rendant compte de jugements globaux. In *European working group on MCDA, bochum, germany*.

Doumpos, M., & Zopounidis, C. (2002). *Multicriteria decision aid classification methods*. Springer.

Fallah Tehrani, A., Cheng, W., & Hüllermeier, E. (2011). Choquistic regression: Generalizing logistic regression using the Choquet integral. In *proceedings EUSFLAT-2011 7th international conference on the European society for fuzzy logic and technology* (pp. 868–875).

Fernández, E., Figueira, J. R., & Navarro, J. (2019). An indirect elicitation method for the parameters of the ELECTRE TRI-nB model using genetic algorithms. *Applied Soft Computing, 77*, 723–733.

Fernández, E., Figueira, J. R., Navarro, J., & Roy, B. (2017). Electre tri-nB: A new multiple criteria ordinal classification method. *European Journal of Operational Research, 263*(1), 214–224.

Figueira, J., Greco, S., Roy, B., & Słowiński, R. (2010). ELECTRE methods: Main features and recent developments. In C. Zopounidis, & P. M. Pardalos (Eds.), *Handbook of multicriteria analysis* (pp. 51–89). Springer.

Figueira, J., Mousseau, V., & Roy, B. (2005). Electre methods. In *Multiple criteria decision analysis: State of the art surveys* (pp. 133–153). Springer.

Furnkranz, J., & Hullermeier, E. (2011). *Preference learning*. Springer. https://doi.org/10.1007/978-3-642-14125-6.

Greco, S., Matarazzo, B., & Slowinski, R. (2001). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research, 129*(1), 1–47.

Greco, S., Mousseau, V., & Slowinski, R. (2010). Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research, 207*(3), 1455–1470.

Herrera-Viedma, E., Herrera, F., Chiclana, F., & Luque, M. (2004). Some issues on consistency of fuzzy preference relations. *European Journal of Operational Research, 154*(1), 98–109.

Jacquet-Lagreze, E., & Siskos, J. (1982). Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. *European Journal of Operational Research, 10*(2), 151–164.

Kadzinski, M., & Ciomek, K. (2021). Active learning strategies for interactive elicitation of assignment examples for threshold-based multiple criteria sorting. *European Journal of Operational Research*.

Kadzinski, M., Ghaderi, M., & Dabrowski, M. (2020). Contingnet preference disaggregation model for multiple criteria sorting problem. *European Journal of Operational Research, 281*(2), 369–387.

Kadzinski, M., & Martyn, M. (2020). Enriched preference modeling and robustness analysis for the Electre Tri-B method. *Annals of Operations Research*, 1–35. https://doi.org/10.1007/s10479-020-03833-z.

Kadziński, M., Greco, S., & Słowiński, R. (2014). Robust ordinal regression for dominance-based rough set approach to multiple criteria sorting. *Information Sciences, 283*, 211–228.

Kadziński, M., Tervonen, T., & Figueira, J. R. (2015). Robust multi-criteria sorting with the outranking preference model and characteristic profiles. *Omega, 55*, 126–140.

Köksalan, M., & Ozpeynirci, S. (2009). An interactive sorting method for additive utility functions. *Computers & Operations Research, 36*(9), 2565–2572.

Labreuche, C. (2011). A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence, 175*(7), 1410–1448.

Leroy, A., Mousseau, V., & Pirlot, M. (2011). Learning the parameters of a multiple criteria sorting method. In *International conference on algorithmic decision theory* (pp. 219–233). Springer.

Liu, J., Kadzinski, M., Liao, X., & Mao, X. (2021). Data-driven preference learning methods for value-driven multiple criteria sorting with interacting criteria. *INFORMS Journal of Computing, 33*(2), 586–606.

Liu, J., Liao, X., Kadziński, M., & Słowiński, R. (2019). Preference disaggregation within the regularization framework for sorting problems with multiple potentially non-monotonic criteria. *European Journal of Operational Research, 276*(3), 1071–1089.

Liu, J., Liao, X., Mao, X., Wang, Y., & Kadzinski, M. (2020). A preference learning framework for multiple criteria sorting with diverse additive value models and valued assignment examples. *European Journal of Operational Research, 286*(3), 963–985.

Marichal, J.-L., Meyer, P., & Roubens, M. (2005). Sorting multi-attribute alternatives: The TOMASO method. *Computers & Operations Research, 32*(4), 861–877.

Minoungou, P., Mousseau, V., Ouerdane, W., & Scotton, P. (2020). Learning an MR–Sort model from data with latent criteria preference directions. In *Proceeding of the 2nd wokshop from multiple criteria Decision aid to Preference Learning (DA2PL)*.

Mousseau, V., Dias, L. C., Figueira, J., Gomes, C., & Clímaco, J. N. (2003). Resolving inconsistencies among constraints on the parameters of an MCDA model. *European Journal of Operational Research, 147*(1), 72–93.

Mousseau, V., & Słowiński, R. (1998). Inferring an ELECTRE TRI model from assignment examples. *Journal of Global Optimization, 12*(2), 157–174.

Perny, P. (1998). Multicriteria filtering methods based on concordance and non-discordance principles. *Annals of Operations Research, 80*, 137–165.

Roy, B. (1991). The outranking approach and the foundations of Electre methods. *Theory and Decision, 31*(1), 49–73.

Roy, B. (1996). *Multicriteria methodology for decision aiding*. Dordrecht: Kluwer Academic.

Rudin, C., & Ertekin, S. (2018). Learning customized and optimized lists of rules with mathematical programming. *Mathematical Programming Computation, 10*, 659–702.

Siskos, Y., Grigoroudis, E., & Matsatsinis, N. (2016). UTA methods. In S. Greco, M. Ehrgott, & J. Figueira (Eds.), *Multiple criteria decision analysis*. In *International Series in OR/MS* (pp. 315–362). Springer.

Sobrie, O. (2016). *Learning preferences with multiple-criteria models*. Université de Mons (Faculté Polytechnique) and Université Paris-Saclay (CentraleSupélec) Ph.D. thesis..

Sobrie, O., Lazouni, M. A., Mahmoudi, S., Mousseau, V., & Pirlot, M. (2016). A new decision support model for preanesthetic evaluation. *Computer Methods and Programs in Biomedicine, 133*, 183–193.

Sobrie, O., Mousseau, V., & Pirlot, M. (2013). Learning a majority rule model from large sets of assignment examples. In P. Perny, M. Pirlot, & A. Tsoukiás (Eds.), *Algorithmic decision theory*. In *Lecture Notes in Artificial Intelligence: 8176* (pp. 336–350). Springer.

Sobrie, O., Mousseau, V., & Pirlot, M. (2015). Learning the parameters of a non compensatory sorting model. In T. Walsh (Ed.), *Algorithmic decision theory*. In *Lecture Notes in Artificial Intelligence: 9346* (pp. 153–170). Lexington, KY, USA: Springer.

Sobrie, O., Mousseau, V., & Pirlot, M. (2019). Learning monotone preferences using a majority rule sorting model. *International Transactions in Operational Research, 26*(5), 1786–1809.

Sokolovska, N., Chevaleyre, Y., & Zucker, J. D. (2018). A provable algorithm for learning interpretable scoring systems. In A. J. Storkey, & F. Pérez-Cruz (Eds.), *International conference on artificial intelligence and statistics, AISTATS 2018* (pp. 566–574).

Tervonen, T., Figueira, J., Lahdelma, R., Almeida Dias, J., & Salminen, P. (2009). A stochastic method for robustness analysis in sorting problems. *European Journal of Operational Research, 192*(1), 236–242.

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning, 102*(3), 349–391.

Zheng, J., Metchebon Takougang, S., Mousseau, V., & Pirlot, M. (2014). Learning criteria weights of an optimistic Electre Tri sorting rule. *Computers & Operations Research, 49*, 28–40.

Preference elicitation for a ranking method based on multiple reference profiles

Alexandru-Liviu Olteanu[1] [ID] · **Khaled Belahcene**[4] · **Vincent Mousseau** [2] [ID] ·
Wassila Ouerdane[2] [ID] · **Antoine Rolland**[3] · **Jun Zheng**[2]

Abstract

Multiple criteria decision aid methodologies support decision makers (DM) facing decisions involving conflicting objectives. DM's preferences should be captured to provide meaningful recommendations. Preference elicitation aims at incorporating DM's preferences in decision models. We propose a new preference elicitation tool for a ranking model based on reference points (RMP—Ranking with Multiple Profiles). Our methodology infers an RMP model from a list of pairwise comparisons provided by the DM. The inference algorithm makes use of a Mixed Integer mathematical programming formulation. We prove the applicability by performing extensive numerical experiments on datasets whose size corresponds to real-world problem.

Keywords Multi-criteria decision aiding · Preference elicitation · Ranking problem · Reference profiles

Mathematics Subject Classification 90B50 · 90C11 · 90C29 · 91B08

1 Introduction

In the field of Multiple Criteria Decision Aiding (MCDA), real world decision problems can be modeled using mainly three classes of problems formulations (see for instance Roy 1996): choice, ranking and sorting. Choice refers to the selection of

Vincent Mousseau
vincent.mousseau@centralesupelec.fr

Wassila Ouerdane
http://wassilaouerdane.github.io

1 Lab-STICC, CNRS, Université de Bretagne Sud, Lorient, France

2 MICS, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

3 ERIC , Université de Lyon 2, Lyon EA 3083, France

4 Heudiasyc, Université de Technologie de Compiègne, Compiègne, France

the best alternative(s), ranking seeks to order all the alternatives from the best one to the worst, whereas sorting aims at assigning each alternative to one of the predefined ordered categories. In this paper, we consider the ranking problem.

Several aggregation methods have been proposed in the literature to rank a set of alternatives (see for instance Figueira et al. 2005; Keeney and Raiffa 1976). In this paper, we are interested in a recently proposed method based on outranking relations, called *Ranking based on Multiple reference Profiles* (RMP) (Rolland 2013; Bouyssou and Marchant 2013), which provides a ranking of alternatives by comparing alternatives to a set of reference profiles. More precisely, we focus on a specific case of this RMP ranking model in which the importance of criteria is represented by additive weights. The RMP method is based on pairwise comparisons, but instead of directly comparing alternatives one to each other, it rather compares each alternative to a set of predefined external reference profiles. The idea is to construct a preference relation on the set of alternatives based on the way each alternative compares with the specified reference profiles.

When used in context with a specific Decision Maker (DM), the RMP ranking model should be tuned so that it accurately reflects the DM viewpoints. Preference elicitation is the process by which an analyst and a decision maker interact in order to set the values for the preference parameters. The direct elicitation approach requires the DM to give explicitly numerical values for the model parameters, whereas the indirect approach uses holistic information provided by the decision maker in order to infer the model parameters (see e.g. Jacquet-Lagrèze and Siskos 2001; Mousseau and Pirlot 2015). The direct elicitation approach is generally considered too difficult to apply in practice, as the DM has no clear understanding of the link between the parameters values and the resulting ranking (Bouyssou et al. 2006). The indirect approach, on the other hand, reduces the cognitive effort required from the DM who is asked to express holistic judgments on alternatives only (e.g. pairwise comparisons of alternatives).

In this work we propose an indirect approach to elicit the parameters of the RMP model from pairwise comparisons expressed by the decision maker. We formulate the elicitation algorithm as a mixed linear optimization problem. In this optimization program, the variables are the parameters of the RMP method, the constraints represent the binary comparisons expressed by the decision maker, and the objective function maximizes the number of restored comparisons.

The paper is structured in the following way: Sect. 2 presents an overview of multicriteria ranking methods in general and reference-based methods in particular. Section 3 introduces the reader to the RMP method through a simple example. In Sect. 4 we provide the technical details on the preference elicitation algorithm in order to infer an RMP model. Section 5 provides a numerical analysis of the behavior of the inference algorithm. We end by concluding remarks and perspectives for future work in Sect. 6.

2 Related literature

MCDA methods are generally classified into two families. The first one concerns methods based on multi-attribute value theory (MAVT) (see Keeney and Raiffa 1976),

while the second includes pairwise comparison methods based (so called outranking methods, see Roy 1991). In this paper, we are interested in a ranking method based on the construction of an outranking relation.

In outranking methods, a preference relation called *outranking relation* is built between pairs of alternatives evaluated on multiple criteria. It is defined as a weak preference relation, noted $\succsim$, on the set of alternatives whose meaning is "*at least as good as*". An alternative a outranks another one b, i.e. $a \succsim b$, if there are strong enough arguments to declare that a is at least as good as b, and if there is no essential reason to refute the statement. Outranking methods includes methods like ELECTRE (Roy 1991), PROMETHEE (Brans et al. 1984), TACTIC (Vansnick 1986). The popularity of such methods lies in their ability to deal with ordinal scales, limited input data and to represent non-compensatory preferences (Bouyssou et al. 2006; Figueira et al. 2005).

The RMP ranking method is an outranking method which involves the use of external profiles to rank alternatives. Numerous studies report psychological evidence that decision makers make decisions based on some references, which can be the current status or their expectations (see for more details Knetsch 1989; Tversky and Kahneman 1991; Samuelson and Zeckhauser 1988; Köszegi and Rabin 2006). The use of reference profiles in preference relations has been already studied in the MCDA literature. For instance, several multi-criteria optimization methods are based on the use of an ideal point. The TOPSIS (the Technique for Order of Preference by Similarity to Ideal Solution) method (Hwang et al. 1993) evaluates an alternative by maximizing the distance between the alternative and an anti-ideal point while minimizing the distance to the ideal point. MACBETH (Bana e Costa and Vansnick 1994) uses two fictitious reference levels on criteria ("*good*" and "*neutral*") to support the elicitation of a value based model.

Reference profiles are also used in sorting problems. For instance, the ELECTRE TRI sorting method (Roy 1991; Figueira et al. 2005) compares alternatives to ordered reference profiles which represent the lower and upper bound of categories. The assignment rules of ELECTRE TRI are very similar to the RMP method. However their output differs; RMP provides a weak ranking of the alternatives while ELECTRE TRI produces a partition of alternatives into predefined categories. The result can be more discriminative for RMP than for ELECTRE TRI; indeed, if the assignment of alternatives to categories is regarded as a partial preorder, the number of equivalent classes is limited by the number of categories. For instance, in the case of ELEC-TRE TRI, comparing alternatives to a profile would result in at most two equivalence classes. In the case of RMP, however, comparing pairs of alternatives to a single profile could result in as many as 2^m, with m being the number of criteria. It should also be emphasized that the RMP ranking method shares some characteristics with the Non-Compensatory Sorting method (NCS, an axiomatic variant of ELECTRE TRI, see Bouyssou and Marchant 2007a, b), and MR-Sort a version in which the importance of criteria coalitions are represented by additive weights. Indeed, as NCS and MR-Sort, RMP defines an outranking relation which defines how alternatives compare to external profiles. Moreover, the mathematical programming formulation of the inference of RMP model from preference statements share some common features with the inference of MR-Sort (see Leroy et al. 2011); This is the case in particular with

Eqs. (4)–(6) in Sect. 4.2 which define how alternatives compare to profiles, and how the weights of supporting coalitions are defined.

The RPM method is also very strongly related to an ordinal aggregation method based on reference points proposed in the framework of decision under uncertainty, see (Perny and Rolland 2006). The elicitation method proposed in this paper could be useful in this context to reveal the reference points and subjective likelihood attached to events by the DM.

As pointed by Bouyssou and Marchant (2013), reference based ranking models are strongly linked with discrete Sugeno integral (Sugeno 1974): a ranking model with a single reference point that is a weak order always has a representation using a discrete Sugeno integral. Such observation makes the literature that aims at eliciting Sugeno integrals of strong interest to our work (see e.g. Prade et al. 2009).

RMP is a method which makes use of pairwise comparisons to derive a ranking. It is well known that multicriteria methods which rely on pairwise comparisons to compute a ranking face a structural difficulty: the presence of Condorcet cycles in the outranking relation. Indeed, the preference relation over alternatives resulting from a weighted majority voting of criteria is not necessarily transitive (see Condorcet 1785). This is why most ranking methods that rely on an outranking relation transform this relation into a transitive ranking, using a so-called exploitation procedure (see e.g. Brans et al. 1984; Figueira et al. 2005). To circumvent the issue of Condorcet cycles, the RMP method proceeds in a slightly different way: pairwise comparisons are not used to compare alternatives one to each other, but to compare alternatives to external profiles (as it is done in ELECTRE TRI Figueira et al. 2005). Hence, no outranking relation is built on the set $\mathcal{A}$ of alternatives; the outranking relation $\succsim$ considered in RMP compares alternatives in $\mathcal{A}$ with the reference profiles in $\mathcal{P}$, i.e, $\succsim \subseteq \mathcal{A} \times \mathcal{P} \cup \mathcal{P} \times \mathcal{A}$. As RMP imposes a dominance structure on the profiles (see Sect. 3.2), the relation $\succsim$ will have no cycles.

Apart from RMP which ranks alternatives based on their comparison to external profiles, outranking based ranking methods rank (order) alternatives according to the way each alternative compares to others. For these pairwise comparison methods the presence or absence of an alternative c can impact the relative rank of two other alternatives a and b. In other words, these outranking methods do not fulfill the property of the Independence of Irrelevant Alternative (IIA), see e.g. for ELECTRE III (Figueira et al. 2005; Wang and Triantaphyllou 2008). Note that fulfilling the IIA property (or not) is neither positive nor negative (one can argue that when a is prefered to c, and c is prefered to b, it grants a comparative advantage of a over b, or not).

This observation has however an important implication concerning indirect preference elicitation with outranking based ranking methods. Suppose we want to rank alternatives in a set $\mathcal{A}$ using an outranking based ranking method called M; we would like to infer the parameters values of this ranking method M from a list of pairwise comparisons provided by the decision maker. Let us denote $\mathcal{A}^* \subset \mathcal{A}$ the alternatives involved in these comparisons, and suppose that two of these alternatives $a, b \in \mathcal{A}^*$ are judged by the decision maker such that a is prefered to b. If the IIA property is not fulfilled by M, there is no formal guaranty that the ranking on $\mathcal{A}$ resulting from the use of the method M using the parameters inferred from the pairwise comparisons of alternatives in $\mathcal{A}^*$ will rank a better than b. Indeed b could be ranked better than

Table 1 Data involved in the illustrative example

	Price (k€)	Confidence in the brand ([0, 100])	Consumption (lit./100 km)	Acceleration (s)
x	18	95	9	24
y	16	66	6	32
z	13	25	6	22
p^1	20	50	10	30
p^2	12	75	7	25
Weight	0.25	0.25	0.25	0.25

a, and this would be difficult to understand from the DM's perspective. This is why it is usually difficult to use outranking based ranking method using indirect preference elicitation. A small example illustrating such difficulty is provided in the Appendix.

The RMP method has a unique advantage over the other outranking based ranking methods, as it is, up to our knowledge, the only outranking based ranking method which fulfills the IIA property. The RMP ranking method can therefore be used meaningfully be used in an indirect elicitation perspective. In what follows, we propose a preference elicitation algorithm to learn the parameters of the RMP method from pairwise comparisons provided by the decision maker.

3 RMP: a ranking model based on multiple profiles

3.1 Illustrative example

In order to provide an overview of how the RMP ranking method proceeds, we provide a small didactic example. Let us consider a decision problem in which cars should be ranked, based on their attractiveness from a buyer's perspective. For the sake of the example, we consider three cars: x, y and z. Each car is evaluated on four criteria: the *price* (in k€, to be minimized), the *confidence in the brand* ([0, 100] scale, the greater the better), the fuel *consumption* (liters per 100 km, to be minimized), and *acceleration* (time in seconds to accelerate from 0 to 100 km/h, to be minimized). The performance of cars are presented in Table 1.

In order to model the judgment of the decision-maker, the RMP method makes use of the following preference parameters: (1) reference profiles, (2) a lexicographic order on these profiles and (3) criteria weights.

In our example, we use two reference profiles denoted with p^1 and p^2 (which are vectors of evaluations), such that p^2_j is better that p^1_j on each criterion j. The values of these two profiles are provided in Table 1. The dominance structure on these two profiles (p^2 dominates p^1) allows to define, on each criterion, three segments on the evaluation scales: better than p^2 (which can be interpreted as *"good"*), between p^1 and p^2 (which can be interpreted as *"intermediate or fair"*), and worse than p^1 (which can be interpreted as *"insufficient"*).
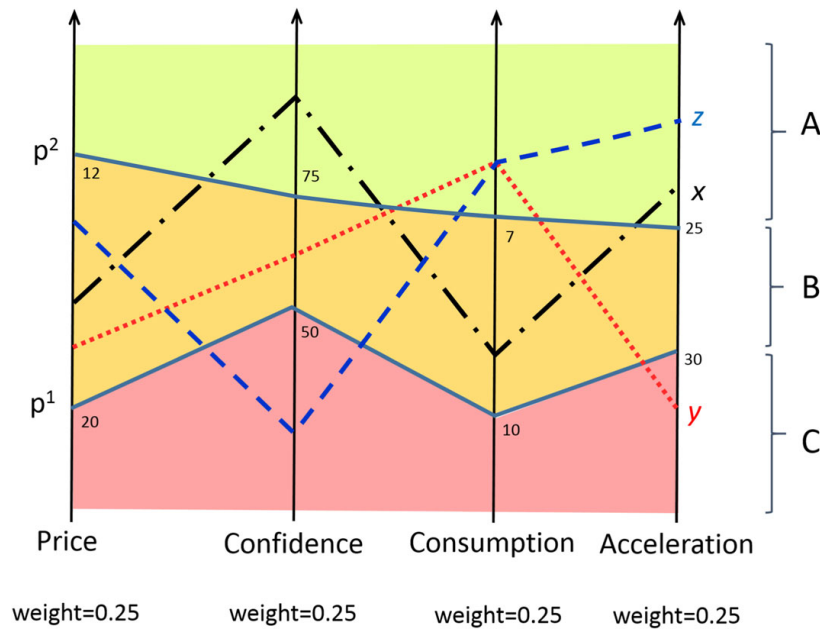
Fig. 1 Graphical interpretation of Table 1

Table 2 Results of the encoding procedure for the illustrative example

	Price	Confidence in the brand	Consumption	Acceleration
x	*B*	*A*	*B*	*A*
y	*B*	*B*	*A*	*C*
z	*B*	*C*	*A*	*A*

In other terms, the reference profiles specify an ordered encoding for each criterion defined by three ordered intervals of performances (A, B, and C) as illustrated in Fig. 1, such that:

A performances better than p^2 on each criterion are denoted A,
B performances between p^1 and p^2 on each criterion are denoted B,
C performances worse than p^1 on each criterion are denoted C.

The RMP method ranks cars based on how they compare to profiles p^1 and p^2. Table 2 shows the encoding of the three cars considered in our example. In addition, a lexicographic order is considered among the reference profiles; this order defines the order by which each car is compared to the profiles. In our case, this order can either be "*compare cars to p^1 then to p^2*" or "*compare cars to p^2 then p^1*". We consider in this example the first one ("*compare cars to p^1 then to p^2*").

To compute a ranking, alternatives are not compared one to each other but each one is compared to the reference profiles. First, alternatives are compared to the first profile in the lexicographic order (here p^1). Considering two alternatives a and b, a is preferred to b, denoted ($a \succ b$), if the number[1] of criteria for which alternative a is

---

[1] In this example, as criteria are equally weighted, we just count the number of criteria, but they could be weighted differently.

evaluated A or B (i.e. better than p^1) is greater than the number of criteria for which alternative b is evaluated A or B. If a and b cannot be distinguished with respect to their comparison to p^1, then a and b are compared to p^2 (the second reference profile in the lexicographic order). If the number of criteria for which alternative a is evaluated A (i.e. better than p^2) is greater than the number of criteria for which alternative is evaluated A, then a is prefered to b, otherwise a and b are indifferent. In our example, we have thus the following:

- Car x is better than car y because, x is evaluated A or B on all criteria while y is evaluated A or B on three criteria only (x compares better to p^1 than y does).
- Car x is better than car z because, x has evaluations A or B on all criteria while z has evaluations A or B on three criteria only (x compares better to p^1 than z does).
- Car z is better than car y because z and y are both evaluated A or B on three criteria (they compare equally to p^1), but z is evaluated A on two criteria while y is evaluated A once only (z compares better to p^2 than y does).

The final ranking is thus: x is the best car, followed by z and then y.

3.2 The RMP ranking method

Let us consider a finite set of alternatives $\mathcal{A}$ evaluated on m criteria. We denote $M = \{1, 2, \ldots, j, \ldots, m\}$ the set of criteria indices, while a_j denotes the evaluation of alternative $a \in \mathcal{A}$ on criterion j (in what follows we will consider, without loss of generality, that preferences increase with the evaluation on each criterion, i.e. the greater the better). Thus, $X = \prod_{j \in M} X_j$ denotes the Cartesian product of evaluations scales X_j. The RMP method makes use of three different types of parameters:

- $\mathcal{P} = \{p^h, h = 1, \ldots, k\}$ a set of k reference profiles, with $p^h = \{p_1^h, \ldots, p_j^h, \ldots, p_m^h\}$, where p_j^h denotes the evaluation of profile p^h on criterion j; we pose without loss of generality[2] a dominance structure on the set of profiles, i.e, $p_j^h \leq p_j^{h+1}$, $\forall h = 1 \ldots k - 1, \forall j \in M$.
- σ, a lexicographic order on the reference profiles, i.e., a permutation on $\{1, \ldots, k\}$. Note that the lexicographic order σ can be any total order on profiles.
- criteria weights $w_1, w_2, \ldots, w_m$, where $w_j \geq 0$ and $\sum_{j \in M} w_j = 1$

RMP proceeds by using a three-step procedure:

1. compute $C(a, p^h) = \{j \in M : a_j \geq p_j^h\}$ with $a \in \mathcal{A}$, $h = 1, \ldots, k$, the set of criteria on which alternative a is at least as good as profile p^h.
2. compare alternatives one to each other in order to define k preference relations $\succsim_{p^h}$, relative to each reference profile such that $a \succsim_{p^h} b$ iff $\sum_{j \in C(a, p^h)} w_j \geq \sum_{j \in C(b, p^h)} w_j$. In other words, $a \succsim_{p^h} b$ holds when a compares at least as good as to p^h than b does. We will denote $\succ_{p^h}$ ($\sim_{p^h}$, respectively) the asymmetric part (the symmetric part, respectively) of the relation $\succsim_{p^h}$.

---

[2] for any RMP model, there exists an equivalent RMP model with a dominance structure on profiles, see (Rolland 2013).

3. rank two alternatives $a, b \in \mathcal{A}$ by sequentially considering the relations $\succsim_{p^{\sigma(1)}}$, $\succsim_{p^{\sigma(2)}}, \ldots, \succsim_{p^{\sigma(k)}}$ (according to the lexicographic order σ); a is prefered to b if $a \succ_{p^{\sigma(1)}} b$, or if $a \sim_{p^{\sigma(1)}} b$ and $a \succ_{p^{\sigma(2)}} b$, or …Hence, a and b are indifferent iff $a \sim_{p^{\sigma(h)}} b$, for all $h = 1, \ldots, k$.

4 The preference elicitation algorithm

In order to set the parameters of an aggregation method, it is necessary to interact with the decision maker, so as to integrate her preferences. A first approach (referred to as direct elicitation in the literature) assumes that the DM understands very well the model and is at ease with expressing the values of its parameters. However, such an approach is not always recommended as the DM does not usually have a clear understanding of the semantics attached to the preference parameters. Therefore, the literature frequently proposes indirect elicitation (see e.g. Mousseau and Slowiński 1998; Leroy et al. 2011), in which the decision maker expresses holistic preferences (i.e. pairwise comparisons on real or fictitious alternatives) from which the values of the parameters are inferred.

We propose to infer, from pairwise comparisons expressed by the DM, the parameters of an RMP model involving:

- the k reference profiles $\mathcal{P} = \{p^h, h = 1, \ldots, k\}$, with $p^h = \{p_1^h, \ldots, p_j^h, \ldots p_m^h\}$;
- the criteria weights w_j, $j \in M$, where $w_j \geq 0$ and $\sum_{j \in M} w_j = 1$;
- the lexicographic order on reference profiles σ.

4.1 Principle

We propose an indirect elicitation procedure for the RMP model, in which the decision maker provides a list $\mathcal{BC}$ of binary comparisons of alternatives (a partial ranking), from which the RMP preference parameters (weights, reference profiles, and the lexicographic order on reference profiles) are inferred. More precisely, two sets are considered $\mathcal{BC}_\succ$ and $\mathcal{BC}_\sim$, such that $\mathcal{BC} = \mathcal{BC}_\succ \cup \mathcal{BC}_\sim$ where $\mathcal{BC}_\succ$ represents the pairs (a, b) for which the decision maker stated that a is preferred to b, while $\mathcal{BC}_\sim$ includes the pairs which are indifferent. We will denote A^* the set of alternatives involved in $\mathcal{BC}$.

With a given number of profiles k, we examine all the $k!$ possible lexicographic orders[3] to identify the RMP model that best matches $\mathcal{BC}$. Formally, this lexicographic order on the profiles is a parameter of the model and should be elicited. If the decision maker can directly provide this order it can be obviously easily taken into account.

Thus, given an order on reference profiles (i.e., for a given lexicographic order σ), determining whether an RMP model fulfilling the preference relations in $\mathcal{BC}_\succ$ and the indifference relations in $\mathcal{BC}_\sim$ amounts to solve a Mixed Integer Program (MIP) which will output the set of criteria weights and the evaluations of the reference profiles. The formulation of this MIP is presented below.

---

[3] In the RMP ranking model, the number of reference profiles is usually limited to 3 or 4. The analysis of 3! (or 4!) orders on profiles is not computationally prohibitive.

Table 3 Parameters of the mathematical model

A^*	The set of alternatives (n in total)
M	The set of criteria indices (m in total)
k	The number of reference profiles
G	The alternatives evaluations, or performance table, given as a matrix of size $n \times m$ (with $g_{a,j} \in [0, 1]$ containing the evaluation of alternative $a \in A^*$ on criterion $j \in M$)
$\mathcal{BC}$	A set of pairs $(a, b) \in A^* \times A^*$ for which the DM provides his/her preference; $\mathcal{BC} = \mathcal{BC}_\succ \cup \mathcal{BC}_\sim$
$\mathcal{BC}_\succ$	A set of pairs $(a, b) \in A^* \times A^*$ where a is preferred to b by the DM
$\mathcal{BC}_\sim$	A set of pairs $(a, b) \in A^* \times A^*$ where a and b are indifferent to the DM
σ	A vector containing a permutation of $1, \ldots, k$ corresponding to a lexicographic order of the reference profiles
γ	A small positive constant used to model strict inequalities

4.2 Mathematical Program for the elicitation algorithm

In this section, we define a mathematical formulation for learning, from a given set of pairwise comparisons provided by the DM, the profiles p^h, $h = 1 \ldots k$, and weights of criteria w_j, $j = 1 \ldots m$, for a given lexicographic order on profiles.[4] Hence, the parameters that are considered as data in the proposed mathematical model are provided in Table 3 above.

We consider, without loss of generality, that the criteria evaluations scales are defined on the unit interval and that larger values are preferred to smaller ones. In order to apply the proposed model to real problem instances, a simple transformation of the criteria scales is required.

The variables considered in the mathematical program correspond to the RMP parameters that are to be inferred (i.e, criteria weights w_j and reference profiles p^h), and additional "technical" variables which are necessary to formulate the constraints. These variables are listed in Table 4.

Given the previous definitions, we introduce hereafter a Mixed-Integer Linear Programming formulation to infer criteria weights w_j and reference profiles p^h from a set $\mathcal{BC}$ of comparisons provided by the DM:

$$\max \quad \sum_{(a,b) \in \mathcal{BC}} t_{a,b} \tag{1}$$

$$\text{s.t. :}$$

$$\sum_{j=1}^m w_j = 1, \text{ with } w_j \geq \gamma$$
$$\forall j \in M \tag{2}$$

$$1 \geq p_j^{h+1} \geq p_j^h \geq 0$$
$$\forall j \in M, \forall h \in 1, \ldots, k-1 \tag{3}$$

---

[4] If this order is unknown, we solve the mathematical program for each possible order; this is reasonable for RMP models with three (or at most four) profiles, which is a standard use of an RMP model.

Table 4 Variables of the mathematical model

w_j	Continuous	The criteria weights of the RMP model, $\forall j \in M$
p_j^h	Continuous	The performance of the reference profiles of the RMP model $\forall j \in M$, $\forall h \in 1, \ldots, k$
$\delta_{a,j}^h$	Binary	1 if alternative a outranks profile h on criterion j ($g_{a,j} \geq p_j^h$), and 0 otherwise $\forall a \in A^*, \forall j \in M, \forall h \in 1, \ldots, k$
$\omega_{a,j}^h$	Continuous	Equal to w_j if $\delta_{a,j}^h = 1$ and to 0 otherwise, $\forall a \in A^*, \forall j \in M$, $\forall h \in 1, \ldots, k$
$s_{a,b}^h$	Binary	1 if alternative a is preferred or indifferent to alternative b w.r.t. profile h, and 0 otherwise, $\forall (a, b) \in \mathcal{BC}, \forall h \in 1, \ldots, k$
$t_{a,b}$	Binary	1 if the comparison between alternative a and b, as given by the DM, is enforced, and 0 otherwise, $\forall (a, b) \in \mathcal{BC}$

$$g_{a,j} - p_j^h + 1 \geq \delta_{a,j}^h \geq g_{a,j} - p_j^h + \gamma$$
$$\forall a \in A^*, \forall j \in M, \forall h \in 1, \ldots, k \tag{4}$$

$$w_j \geq \omega_{a,j}^h \geq 0$$
$$\forall a \in A^*, \forall j \in M, \forall h \in 1, \ldots, k \tag{5}$$

$$\delta_{a,j}^j \geq \omega_{a,j}^h \geq \delta_{a,j}^j + w_j - 1$$
$$\forall a \in A^*, \forall j \in M, \forall h \in 1, \ldots, k \tag{6}$$

$$\sum_{j=1}^m \omega_{a,j}^{\sigma(h)} \geq \sum_{j=1}^m \omega_{b,j}^{\sigma(h)} + \gamma - \left(1 - s_{a,b}^{\sigma(h)} + s_{a,b}^{\sigma(h-1)}\right) \cdot (1 + \gamma)$$
$$\forall (a, b) \in \mathcal{BC}, \forall h \in 2, \ldots, k - 1 \tag{7}$$

$$\sum_{j=1}^m \omega_{a,j}^{\sigma(h)} \geq \sum_{j=1}^m \omega_{b,j}^{\sigma(h)} - s_{a,b}^{\sigma(h)} - s_{a,b}^{\sigma(h-1)}$$
$$\forall (a, b) \in \mathcal{BC}, \forall h \in 2, \ldots, k - 1 \tag{8}$$

$$\sum_{j=1}^m \omega_{a,j}^{\sigma(h)} \leq \sum_{j=1}^m \omega_{b,j}^{\sigma(h)} + s_{a,b}^{\sigma(h)} + s_{a,b}^{\sigma(h-1)}$$
$$\forall (a, b) \in \mathcal{BC}, \forall h \in 2, \ldots, k - 1 \tag{9}$$

$$\sum_{j=1}^m \omega_{a,j}^{\sigma(1)} \geq \sum_{j=1}^m \omega_{b,j}^{\sigma(1)} + \gamma - \left(2 - s_{a,b}^{\sigma(1)} - t_{a,b}\right) \cdot (1 + \gamma)$$
$$\forall (a, b) \in \mathcal{BC} \tag{10}$$

$$\sum_{j=1}^m \omega_{a,j}^{\sigma(1)} \geq \sum_{j=1}^m \omega_{b,j}^{\sigma(1)} - s_{a,b}^{\sigma(1)} - \left(1 - t_{a,b}\right)$$
$$\forall (a, b) \in \mathcal{BC} \tag{11}$$

$$\sum_{j=1}^m \omega_{a,j}^{\sigma(1)} \leq \sum_{j=1}^m \omega_{b,j}^{\sigma(1)} + s_{a,b}^{\sigma(1)} + \left(1 - t_{a,b}\right)$$
$$\forall (a, b) \in \mathcal{BC} \tag{12}$$

$$\sum_{j=1}^{m} \omega_{a,j}^{\sigma(k)} \geq \sum_{j=1}^{m} \omega_{b,j}^{\sigma(k)} + \gamma - s_{a,b}^{\sigma(k-1)} \cdot (1 + \gamma)$$

$$\forall (a, b) \in \mathcal{BC} \tag{13}$$

$$\sum_{j=1}^{m} \omega_{a,j}^{h} \leq \sum_{j=1}^{m} \omega_{b,j}^{h} + (1 - t_{a,b})$$

$$\forall (a, b) \in \mathcal{BC}_{\sim}, \forall h \in 1, \ldots, k \tag{14}$$

$$\sum_{j=1}^{m} \omega_{a,j}^{h} \geq \sum_{j=1}^{m} \omega_{b,j}^{h} + (1 - t_{a,b})$$

$$\forall (a, b) \in \mathcal{BC}_{\sim}, \forall h \in 1, \ldots, k \tag{15}$$

The objective function in (1) seeks to maximize the number of pairwise comparisons (preference and indifference), provided by the DM, that are correctly reproduced by the inferred RMP model.

Constraint (2) is needed to normalize criteria weights. A non-zero lower-bound is added in order to have all the criteria play a role in the pairwise comparisons of alternatives.

Constraint (3) is used to bound the profiles evaluations in the unit interval and ensure the dominance structure on the set of profiles. We assume here, without loss of generality, that the all criteria scales are in the unit interval and that larger evaluations are preferred to lower ones.

Constraints (4) model $\delta_{a,j}^{h}$ variables such that these binary variables are equal to 1 if the evaluation of alternative a on criterion j is greater or equal than the evaluation of profile p^h ($g_{a,j} \geq p_j^h$), and 0 otherwise.

Constraints (5)–(6) model $\omega_{a,j}^{h}$ variables as the minimum value between w_j and $\delta_{a,j}^{h}$. Hence, when $\delta_{a,j}^{h} = 1$ then $\omega_{a,j}^{h} = w_j$ and when $\delta_{a,j}^{h} = 0$ then $\omega_{a,j}^{h} = 0$. These variable will be used in order to compute the sum of the weights of criteria on which an alternative is at least as good as a reference profile: $\sum_{j=1}^{m} \omega_{b,j}^{h}$.

Constraints (7), (8) and (9) model the preference of a over b for each $(a, b) \in \mathcal{BC}$, based on how a and b compare to reference profile $p^{\sigma(h)}$. We make use of binary variables $s_{a,b}^{h}$ which are equal to 1 if alternative a is preferred to alternative b w.r.t. profile $p^{\sigma h}$, and 0 when alternative a is considered indifferent to alternative b w.r.t. this same profile. The reference profile is indexed here using the lexicographic order as we consider how a and b compare to the previous profile in the lexicographic order.

Constraint (7) models a strict preference in favor of alternative a over alternative b w.r.t. profile $p^{\sigma(h)}$. Constraints (8) and (9) model an indifference between a and b w.r.t. profile $p^{\sigma(h)}$. In the case of a strict preference, the sum of the criteria weights on which alternative a is as least as good as profile $p^{\sigma(h)}$ needs to be strictly greater than the sum of the criteria weights on which alternative b is as least as good as profile $p^{\sigma(h)}$. If this constraint can be fulfilled, than the binary variable $s_{a,b}^{\sigma(h)}$ will be 1. Otherwise, $s_{a,b}^{\sigma(h)}$ will be 0 and result in relaxing the constraint.

If a strict preference between alternatives a and b cannot be modeled using profile $p^{\sigma(h)}$, then a and b need to be indifferent w.r.t. this profile. Constraints (8) and (9)

model this statement, and the binary variable $s_{a,b}^{\sigma(h)}$ is used to relax these constraints when constraint (7) is able to model the strict preference relation.

When a strict preference relation is achieved w.r.t to a reference profile $p^{\sigma(h)}$, we relax all constraints on the reference profiles that follow it in the lexicographic order. For this reason, we include the $s_{a,b}^{\sigma(h-1)}$ terms in these three constraints, so that if a strict preference is modeled using a prior reference profile, all three constraints are relaxed.

We consider the special case concerning the first profile in the lexicographic order, in constraints (10), (11) and (12). These constraints are identical to the previous three, except that the last terms correspond to the binary variables $t_{a,b}$ which are used to enforce (when $t_{a,b} = 1$) or relax (when $t_{a,b} = 0$) the global strict preference relation between alternatives a and b. By fixing $t_{a,b} = 1$, the constraints for the first profile enforce either that alternative a is strictly preferred to alternative b w.r.t. this profile, or that alternative a is indifferent to alternative b w.r.t. it. In case alternative a is indifferent to alternative b w.r.t. the first profile in the lexicographic order, than $s_{a,b}^{\sigma(1)} = 1$ therefore the second profile is constrained to model either a strict preference between alternative a and alternative b or an indifference, and so on. Constraint (13) is used to stop this propagation when considering the last profile in the lexicographic order and therefore enforcing that a strict preference between alternatives a and b is modeled.

Finally, constraints (14) and (15) model the indifference relation between all pairs of alternatives $(a, b) \in \mathcal{BC}_{\sim}$. For this, alternatives a and b need to be considered as indifferent for all reference profiles p^h, $h \in 1, \ldots, k$. This means that for each profile p^h, the sum of the weights of the criteria on which a is at least as good as p^h needs to be equal to the sum of the weights of the criteria on which b is at least as good as p^h. The binary variable $t_{a,b}$ is used to relax these constraints (when $t_{a,b} = 0$) if this relation cannot be modelled.

4.3 Integration into an interactive process

The main aim of an elicitation process is to capture the DMs' preferences in order to accurately specify the decision model parameters. It is by definition an interactive process between an analyst and a decision maker. The proposed elicitation method, in this paper, can be naturally embedded in an interactive process with a decision maker. The algorithm can be used in order to provide a first model to be proposed, and thus ask the decision maker if he/she considers the proposed model as acceptable (values of the profiles, weights values, etc.). If the DM expresses a disagreement (for instance: a particular criterion is more important than the others, the value of a profile should be modified,…), this can be easily implemented by a new constraint in the inference program. Thus, the elicitation program integrating the new constraint can be solved to take into account the modification expressed by the decision maker. This makes it possible to progressively refine the model to account for the DM preferences (an example of such an approach is the one is described in Dias et al. (2002)). As the main contribution of the paper is the elicitation method (algorithm), we leave for future work the designing of the interactive process.

5 Numerical analysis

We propose in this section to study the performances of the proposed elicitation algorithm for the RMP model. We begin by describing our experimental protocol. Then, we provide the results concerning (1) the computation time, (2) the ability of the proposed approach to restore the provided binary comparisons, and (3) its ability to handle noisy data.

5.1 Experiment design and implementation details

To test our algorithm we follow the experimental design depicted in Fig. 2. We randomly draw an initial RMP model, denoted with M_o:

- the criteria weights are first randomly generated using the method described in Butler et al. (1997); Zheng et al. (2014),
- the reference profiles are drawn as follows: on each criterion $j \in M$, randomly we generate k_o evaluations in X_j and order them. These ordered evaluations on all criteria are used to specify the k_o profiles, so as to respect the dominance structure on profiles.
- we randomly select a lexicographic order on profiles.

We randomly generate a training set (denoted A_{tr}) of n_{tr} alternatives defined by their evaluations on the m criteria. We then construct a set of n_{bc} pairwise comparisons, by randomly selecting pairs of alternatives from A_{tr} (we discard pairs of alternatives involving dominance); we use M_o in order to extract the preference relations on the selected pairs.

Then we compute M_e, the RMP model that best matches the n_{bc} pairwise comparisons, with a fixed number of profiles k_e using the algorithm proposed in Sect. 4. To appreciate the distance between M_e and M_o, we randomly generate A_{te}, a test set of n_{te} alternatives (A_{te} is constructed in the same way as A_{tr}). A_{te} is used with both the original model and the elicited one in order to construct two rankings of the alternatives. Kendall's rank correlation is then used in order to measure the closeness between the elicited model M_e and the original one M_o.
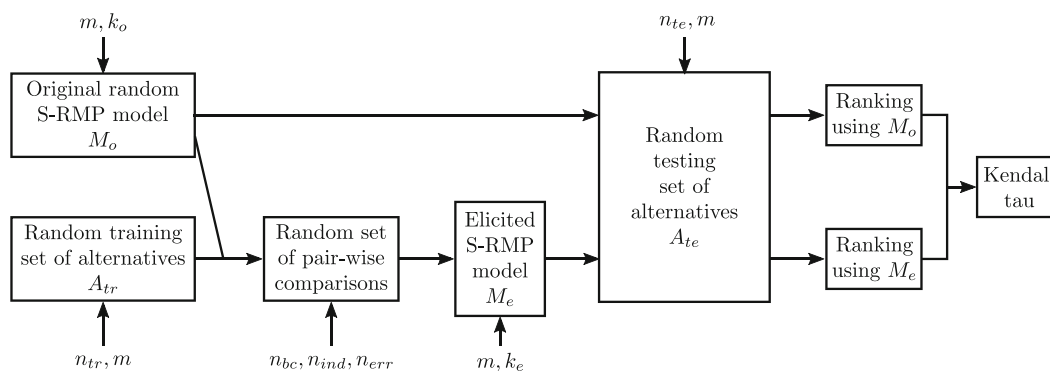
Fig. 2 Design of experiments

We have set the following values for the experiments' parameters (from Fig. 2): $m \in \{3, 5, 7\}$, $n_{tr} \in \{10, 20, \ldots 100\}$, $k_o = 10$, $k_e \in \{1, 2, 3\}$ and $n_{te} = 5000$. We have generated 100 RMP models (M_o) for each combination of values for these parameters. The experiments have been performed using the solver IBM ILOG CPLEX 12.6.3 on an AMD Opteron^{TM} 6176 SE machine with 250 GB RAM and the possibility of launching up to 18 threads in parallel. We have set a 60 min timeout for each computation.

Moreover, when inferring an RMP model, in order to remove any bias caused by the sequence in which the lexicographic orders of profiles were chosen, we have adapted the approach so that multiple parallel executions are launched, one for each lexicographic order. Therefore, when $k = 1$ we launch a single instance of the approach, allowing CPLEX to reach a parallelism of 18, when $k = 2$ we launch two instances with a parallelism of 9 each, and when $k = 3$ we launch six instances with a parallelism of 3 each. In this way, all executions of this algorithm, regardless of the sought number of profiles, will have access to the same amount of resources and the final result will not be biased by the order in which the lexicographic orders have been chosen.

5.2 Experiments results

5.2.1 Computing time

Our first experiment aims to analyze the computation time of the proposed algorithm. Figure 3 depicts the average execution times and standard deviations for the problem instances that were solved within a one hour time limit.

We observe that the execution time increases with the number of considered binary comparisons at an exponential rate. The computation time also increases significantly with the number of profiles.

Note that the exponential trend of the computing time seems to "weaken" for more than 70 comparisons. This may be due to the one hour limit imposed for computing M_e; indeed, this timeout occurs more often for large instances. Finally, the differences in execution time when computing RMP models with one, two or three profiles ($k_e = 1, 2,$ or 3) seem to reduce when more criteria are considered ($m = 7$).

We can conclude that, for data sets whose size corresponds to real instances (up to 100 comparisons, 7 criteria), the computation time is compatible with a working session mode (see for instance Ferretti et al. 2018) in which preference statements are collected from the decision maker, and results are shown after at most 20–30 mn computation. It is however difficult to envisage an interactive trial and error mode with datasets of real world size.

5.2.2 Inferring from noise free data

This section concerns the ability of elicited RMP models to restore a set of noise free binary comparisons. We study the proportion of comparisons restored when the number of comparisons, number of criteria and number of profiles vary.
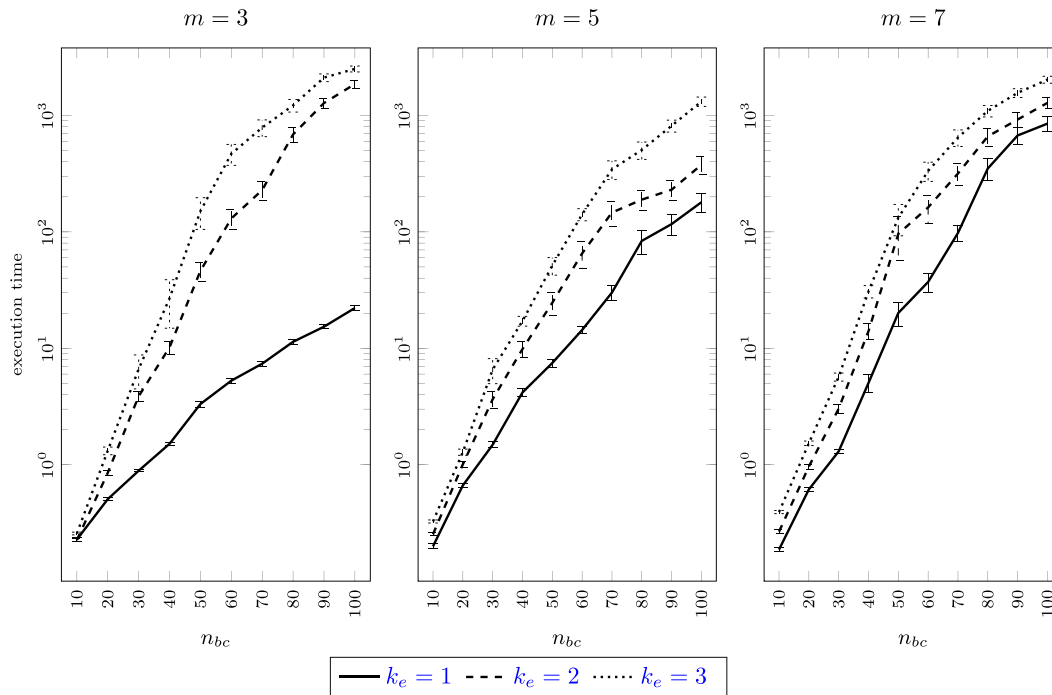
Fig. 3 Average execution times (log scale) wrt number of criteria (m), number of profiles (k_e), and number of comparisons (n_{bc})

- **Flexibility of the RMP model: ability of an RMP model with $k = 1 \dots 3$ profiles to represent a training set generated by an RMP with 10 profiles.** Figure 4 depicts the mean value and standard deviation for the input data restoration rate, i.e, the proportion of the input pairwise comparisons that are correctly restored by the inferred model M_e (as compared to the ground truth M_o).
  Note that, in our experiment, the binary comparisons are generated with an RMP model M_o using $k_o = 10$ profiles whereas the inferred models M_e use a number of profiles k_e varying from 1 to 3. Therefore, there is no guarantee that the inferred models fully restore the set of pairwise comparisons. However, the results depicted in Fig. 4 prove the RMP model to be highly flexible. For instance, with 5 criteria it was always possible in the experiments to restore 100 comparisons generated with a 10 profiles RMP model, with a model using only 2 profiles.
  With 7 criteria ($m = 7$), and for small numbers of binary comparisons ($n_{bc} \leq 40$), it is possible to restore all comparisons generated from M_o even with S-RPM models with a RMP model with a single profile ($k_e = 1$). Decreasing the number of criteria leads to a less flexible model M_e and therefore to a reduced restoration rate. Similarly, increasing the number of profiles ($k_e = 2$ or even 3) improves the flexibility of M_e, and consequently improves the restoration rate.
- **Ability of the inferred model M_e to restore the original one M_o.** In this situation, we test the ability of the elicitation algorithm to compute a model M_e which is as close as possible to the original one M_o. Since an RMP model contains multiple sets of parameters, comparing two models given by two different sets of these parameters can prove rather difficult. We overcome this problem by generating a large test set of alternatives A_{te} and computing the Kendall Tau rank correlation
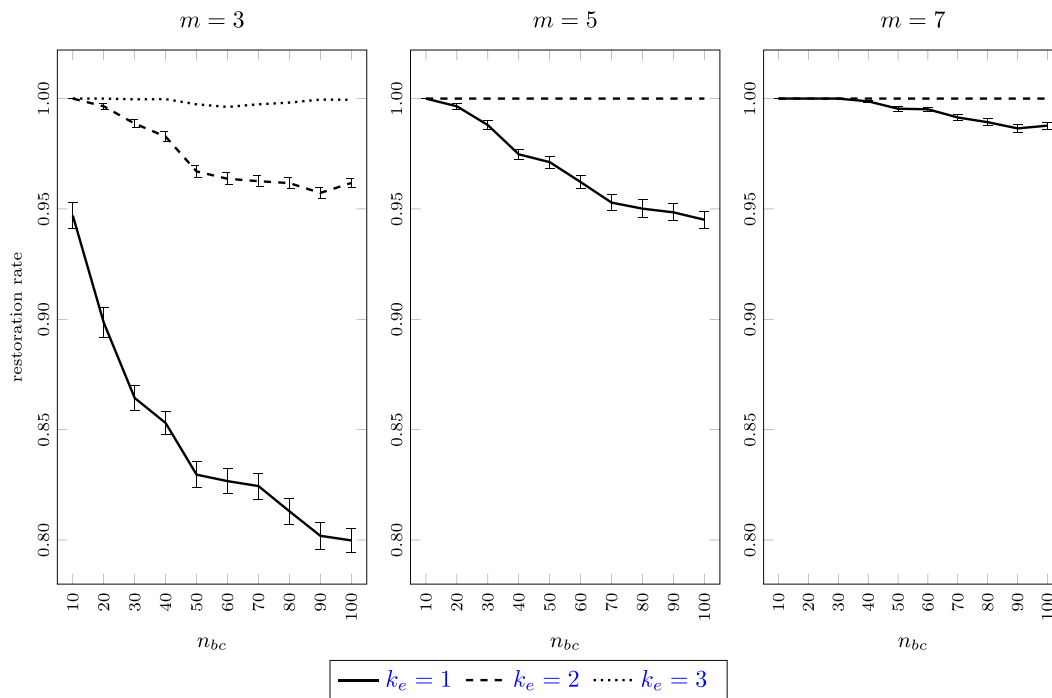
Fig. 4 Proportion of restored binary comparisons on the training data

between the rankings obtained using M_e and M_o. This serves as a proxy for the similarity of the two models, the higher the rank correlation meaning the closer the two models are to each other.

Figure 5 depicts the mean value and the standard deviation of the Kendall Tau between M_e and M_o ranks for $m = 3,\ 5$ and 7, and $k_e = 1,\ 2$ and 3. The experimental results show an expected trend in which increasing the number of input comparisons results in an improvement in the Kendall Tau. The increasing curves of the Kendall Tau values as the number of comparisons increases seem to reach an asymptote: for example, beyond 50 comparisons, in the the case $k_e = 1,\ m = 3$, the Kendall Tau reaches a "plateau" (~ 0.75), and in the the case $k_e = 1,\ m = 5$, the Kendall Tau reaches a "plateau" beyond 80 comparisons (~ 0.85).

With 2 or 3 profiles, we do not observe the asymptote, but we can expect it for a higher number of comparisons. Similarly, as the number of criteria increases, the model gains flexibility, and we observe that more comparisons are required to faithfully elicit the model.

There are configurations (e.g. $m = 7$) in which one would need more comparisons to reach an asymptote and accurately assess the model. This would be computationally costly. However, one should keep in mind that the comparisons are chosen randomly, without any consideration concerning the amount of information provided. To overcome such difficulty, it would be wise to follow an "active learning" approach in which comparisons are carefully selected to provide effective information.
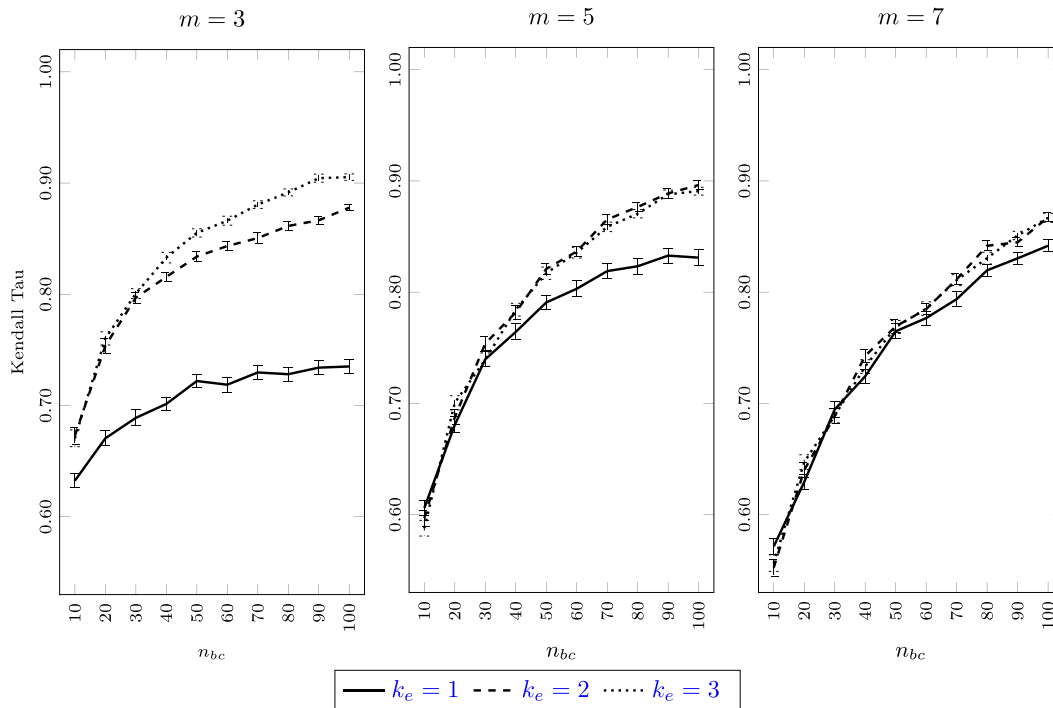
Fig. 5 Kendal Tau between the M_0 and M_e rankings on the test set

Note that when generating pairwise comparisons using an RMP model using 1 profile (2 or 3 profiles respectively) and learning from this comparisons an RMP model with 1 profile (with 2 or 3 profiles respectively), it is obvious that the ability to restore the original model should be higher for an RMP model having less parameters and thus for smaller values of k. However, we do not observe such a trend in Fig. 5. This is due to the fact we generate pairwise comparisons with an RMP model with 10 profiles and learn from these comparisons an RMP model using 1, 2 or 3 profiles. In the results, the main observed phenomenon corresponds to the fact that it is not possible to restore much of the learning set (see Fig. 4), and hence the generalization ability is impoverished.

5.2.3 Results on the ability of the inferred model to restore the original one with noisy data

In the previous experiments, input data was assumed to be noise free. In what follows, we study the effects of introducing a percentage of errors within the input comparisons. By errors, we mean that, after generating a set of binary comparisons with the original model (M_0), we reverse the preference between a proportion of pairs of alternatives. More precisely, we study the situations in which we introduce 5%, 10%, and 15% of "errors" in the set of pairwise comparisons used to infer the model M_e.

Figure 6 depicts the mean value and the standard deviation of the Kendall Tau between M_e and M_o ranks for $m = 5$, $k_e = 1$, and $m = 5$, $k_e = 2$, with a proportion of 5%, 10%, and 15% of "errors". The situation with 0% errors corresponds to the case presented in Fig. 5 for $m = 5$.
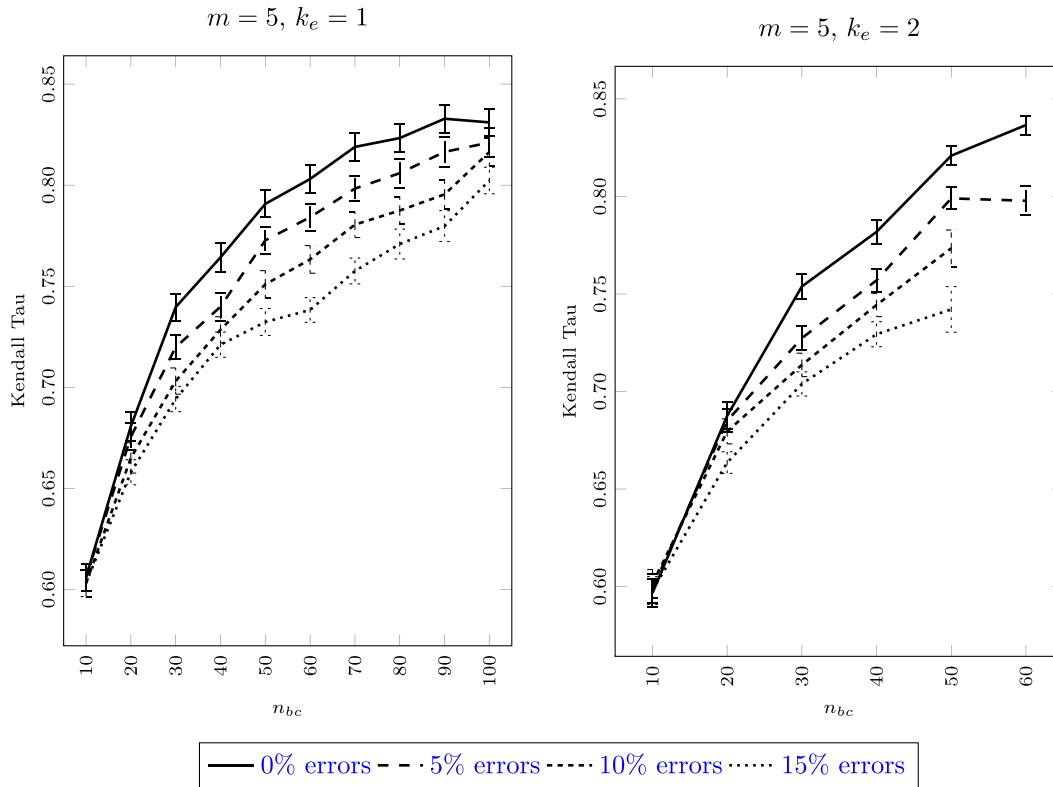
Fig. 6 Kendal Tau between M_0 and M_e rankings on the test set in presence of errors

We observe that, despite the introduction of errors in the input comparisons (event for 15% errors), the algorithm takes advantage of additional comparisons, and the elicited model M_e gets closer to the original one M_o (the Kendall Tau increases). This denotes a positive behavior of the algorithm, even with noisy data.

Obviously, a greater proportion of errors requires a larger number of comparisons to faithfully elicit the model. For instance, we can observe in Fig. 6 that, for $m = 5$ and $k_e = 1$, an average number of 60 noise free comparisons leads to a Kendall Tau equal to ∼0.8; with 5% errors (with 15%, respectively), 80 comparisons (100 comparisons, respectively) are necessary to obtain a similar result.

With noisy data, the computing time increases as compared to noiseless situations. This is illustrated in Table 5 by the percentage of instances that were not solved within a one hour execution time. For $k_e = 1$ (except for $n_{bc} = 100$), all instances were solved in less than 60 mn; however, for $k_e = 2$, starting from $n_{bc} = 60$, more than half of the considered experiments did not provide a solution when errors were included in the binary comparisons.

6 Conclusion and perspectives

In this paper, we propose an indirect approach for the elicitation of the parameters of an RMP model. This approach aims to offer an operational tool to support the use of RMP in real world applications (see e.g. Ferretti et al. 2018). Moreover, the proposed

Table 5 Percentage of instances that did not provide a solution within one hour when considering errors in the binary comparisons for $m = 5$

k_e	Errors (%)	n_{bc} 10	20	30	40	50	60	70	80	90	100
1	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	1
	10	0	0	0	0	0	0	0	0	0	6
	15	0	0	0	0	0	0	0	0	0	19
2	0	0	0	0	0	0	1	8	9	17	19
	5	0	0	0	1	8	55	78	98	100	100
	10	0	0	0	9	50	93	100	100	100	100
	15	0	0	0	18	75	100	100	100	100	100

method has been implemented in R as part of the library of MCDA methods proposed by Bigaret et al. (2017), and is therefore available for use.

For an effective use in real-world applications, computing time can still be an issue for instances of large size, or for situations in which the preference data collected from the decision maker is highly noisy. For such situations, the metaheuristic developed in Liu (2016) could be suitable as it makes it possible to infer in a reasonable computing time an RMP model; such an approach does not however guaranty optimality. Another possibility is to combine, exact and metaheuristics approaches as it was described in Ferretti et al. (2018).

An interesting direction to improve this issue is to express the inference problem as a Boolean Satisfiability Problem (SAT) in order to find a model fully consistent with the learning set (whenever it exists). A first work has already been proposed for multicriteria sorting models, and has proved to be computationally more efficient than optimization approaches (Belahcène et al. 2018). Moreover, additional questions are of interest in relation with this work, for instance empirically exploring the descriptive power of the ordinal ranking model based on reference points, or designing the interactive process to allow the decision maker to give feedback and allow refining the parameters of the model.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

Ethical statement The authors of this paper conform to the Springer Publishing Ethics Statement.

Appendix: Invariance with respect to Irrelevant alternative, the case of pairwise comparison methods

Consider a multiple criteria ranking method in which a weak preference relation $\succsim$ is constructed on the set of alternatives, based on a weighted voting of criteria, and where the ranking is defined computing on $\succsim$ the net flow score of alternatives [5]. More precisely, we consider the relation $\succsim$ defined on $\mathcal{A}$ as follows. for any pair $x, y \in \mathcal{A}$:

$$x \succsim y \Leftrightarrow \sum_{j:x_j \geq y_j} w_j \geq \sum_{j:y_j \geq x_j} w_j$$

where $x = (x_1, ..., x_m)$ and $y = (y_1, ..., y_m)$. The relation $\succsim$ is exploited to rank alternatives using the net flow score $NF(x)$, $x \in \mathcal{A}$: $NF(x) = |f_l(x)| - |f_e(x)|$ where $f_l(x)$ ($f_e(x)$, respectively) represents the leaving flow of x (the entering flow of x, respectively), and is defined by $f_l(x) = \{y \in \mathcal{A} : x \succsim y\}$ ($f_e(x) = \{y \in \mathcal{A} : y \succsim x\}$, respectively).

Consider a small example involving 3 criteria (to be maximized) and 6 alternatives ($\mathcal{A} = \{a, b, c, d, e, f\}$) with the alternatives evaluations described in Table 6.

Suppose that the DM is able to provide preference information concerning $\mathcal{A}^* \subset \mathcal{A}$ a reference set of alternatives $\mathcal{A}^* = \{a, b, c, d\}$ through the form of a ranking on $\mathcal{A}^*$: $a \succ b \succ c \succ d$. Note that the informational content of this ranking boils down to the fact that none of the three criteria is a majority alone (i.e, $w_1 + w_2 > w_3$, $w_1 + w_3 > w_2$ and $w_2 + w_3 > w_1$). Hence, any inference program which would compute the criteria weights from this ranking will find weights compatible with these three inequalities. The computation of the ranking on $\mathcal{A}^*$ using such weights using the net flow score is provided in Table 7, and leads to the ranking: $a \succ b \succ c \succ d$.

Table 6 Small example

	$crit_1$	$crit_2$	$crit_3$
a	2	4	2
b	1	3	5
c	0	5	1
d	5	0	0
e	4	1	4
f	3	2	3

Table 7 Outranking relation on $\mathcal{A}^*$

Γ	a	b	c	d	$|f_l(x)|$
a	1	1	1	1	4
b	0	1	1	1	3
c	0	0	1	1	2
d	0	0	0	1	1
$|f_e(x)|$	1	2	3	4	

Table 8 Outranking relation on $\mathcal{A}$

Γ	a	b	c	d	e	f	$|f_l(x)|$
a	1	1	1	1	0	0	4
b	0	1	1	1	1	1	5
c	0	0	1	1	0	0	2
d	0	0	0	1	0	0	1
e	1	0	1	1	1	1	5
f	1	0	1	1	1	0	4
$|f_e(x)|$	3	2	5	6	3	2	

Suppose now that we want to compute the ranking on the whole set $\mathcal{A}$ (including e and f), based on the weights inferred from the ranking on $\mathcal{A}^*(a \succ b \succ c \succ d)$, i.e., using weights such that $w_1 + w_2 > w_3$, $w_1 + w_3 > w_2$ and $w_2 + w_3 > w_1$. This is provided in Table 8 bellow, and leads to a ranking in which b is ranked first, a is ranked second, then e and f at equal rank, then c then d. It appear that b is ranked better than a, in contradiction with the initially provided preference ranking.

Hence, it appears that, when using such ranking method, the ranking on $\mathcal{A}$ using the weights inferred from the ranking on $\mathcal{A}^*$ (provided by the DM) does not necessarily extend the ranking on $\mathcal{A}^*$. Such observation, makes it difficult to use such ranking method in a disagregation perspective.

References

Bana e Costa CA, Vansnick J -Cl (1994) MACBETH: an interactive path towards the construction of cardinal value functions. Int Trans Oper Res 1:489–500

Belahcène Kh, Labreuche Ch, Maudet N, Mousseau V, Ouerdane W (2018) An efficient SAT formulation for learning multicriteria non-compensatory sorting models. Comput Oper Res 87:58–712

Bigaret S, Hodgett R, Meyer P, Mironova T, Olteanu A-L (2017) Supporting the multi-criteria decision aiding process: R and the MCDA package. EURO J Decis Processes 5(1–4):169–194

Bouyssou D, Perny P (1992) Ranking methods for valued preference relations: A characterization of a method based on leaving and entering flows. Eur J Oper Res 61(1):186–194

Bouyssou D, Marchant T (2007a) An axiomatic approach to noncompensatory sorting methods in MCDM, I: The case of two categories. Eur J Oper Res 178(1):217–245

Bouyssou D, Marchant T (2007b) An axiomatic approach to noncompensatory sorting methods in MCDM, II: More than two categories. Eur J Oper Res 178(1):246–276

Bouyssou D, Marchant T, Pirlot M, Tsoukiàs A, Vincke Ph (2006) Evaluation and decision models with multiple criteria : stepping stones for the analyst

Bouyssou D, Marchant T (2013) Multiattribute preference models with reference points. Eur J Oper Res 229(2):470–481

Brans JP, Maréchal B, Vincke Ph (1984) PROMETHEE: a new family of outranking methods in multicriteria analysis. Oper Res IFORS 84:477–490

Butler J, Jia J, Dyer J (1997) Simulation techniques for the sensitivity analysis of multi-criteria decision models. Eur J Oper Res 103:531–546

Condorcet M (1785) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris

Dias L, Mousseau V, Figueira J, Climaco J (2002) An aggregation/disaggregation approach to obtain robust conclusions with ELECTRE TRI. Eur J Oper Res 138(2):332–348

Ferretti V, Liu J, Mousseau V, Ouerdane W (2018) Reference-based ranking procedure for environmental decision making: Insights from an ex-post analysis. Environ Modell Softw 99:11–24

Figueira J, Mousseau V, Roy B (2005) ELECTRE methods. In: Multiple Criteria Decision Analysis: State of the Art Surveys, pp 133–162. Springer Verlag, New York

Hwang C, Young-Jou L, Ting-Yun L (1993) A new approach for multiple objective decision making. Comput Oper Res 20(8):889–899 Elsevier

Jacquet-Lagrèze E, Siskos Y (2001) Preference disaggregation: 20 years of MCDA experience. Eur J Oper Res 130(2):233–245

Keeney RL, Raiffa H (1976) Decision with multiple objectives: preference and values tradeoffs. Wiley, New York

Knetsch JL (1989) The endowment effect and evidence of nonreversible indifference curves. Am Econ Rev 79(5):1277–1284

Köszegi B, Rabin M (2006) A model of reference-dependent preferences. Q J Econ 121(4):1133–1165

Leroy A, Mousseau V, Pirlot M (2011) Learning the parameters of a multiple criteria sorting method. In: Brafman R, Roberts F, Tsoukiàs A (eds), Algorithmic Decision Theory, LNAI vol. 6992, pp 219–233

Liu J (2016) Preference Elicitation for Multi-Criteria Ranking with Multiple Reference Points. PhD Thesis, Université Paris Saclay

Mousseau V, Pirlot M (2015) Preference elicitation and learning. EUR J Decis Process 3(1–2):1–3

Mousseau V, Slowiński R (1998) Inferring an ELECTRE TRI model from assignment examples. J Global Optim 12(2):157–174

Perny P, Rolland A (2006) Reference-dependent Qualtitative Models for Decision Making under Uncertainty. In: Proceeding of the european conference on artificial intelligence, pp. 422-426

Prade H, Rico A, Serrurier M (2009) Elicitation of Sugeno integrals: a version space learning perspective. foundations of intelligent systems, In: Rauch J, Ras ZW, Berka P, Elomaa T (eds.). Proceedings of the 18th international symposium, ISMS

Rolland A (2013) Reference-based preferences aggregation procedures in multi-criteria decision making. Eur J Oper Res 225(3):479–486

Roy B (1991) The outranking approach and the foundations of ELECTRE methods. Theory and Decision 31:49–73

Roy B (1996) Multicriteria methodology for decision aiding. Kluwer Academic, Dordrecht

Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. J Risk Uncertain 1:7–59

Sugeno M (1974) Theory of Fuzzy Integrals and Its Applications. PhD Thesis, Tokyo Institute of Technology

Tversky A, Kahneman D (1991) Loss aversion in riskless choice: A reference-dependent model. Q J Econ 106(4):1039–1061

Vansnick J-Cl (1986) On the problem of weights in multiple criteria decision making (the noncompensatory approach). Eur J Oper Res 24(2):288–294

Wang X, Triantaphyllou E (2008) Ranking irregularities when evaluating alternatives by using some ELECTRE methods. OMEGA 36(1):45–63

Zheng J, Metchebon Takougang SA, Mousseau V, Pirlot M (2014) Learning criteria weights of an optimistic Electre Tri sorting rule. Comput Oper Res 49:28–40
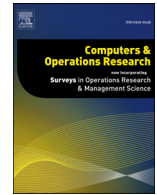
An efficient SAT formulation for learning multiple criteria non-compensatory sorting rules from examples

K. Belahcène [a], C. Labreuche [b], N. Maudet [c], V. Mousseau [a,*], W. Ouerdane [a]

[a] LGI, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
[b] Thales Research & Technology, Palaiseau Cedex 91767, France
[c] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, Paris F-75005, France

ABSTRACT

The literature on Multiple Criteria Decision Analysis (MCDA) proposes several methods in order to sort alternatives evaluated on several attributes into ordered classes. Non-Compensatory Sorting models (NCS) assign alternatives to classes based on the way they compare to multicriteria profiles separating the consecutive classes. Previous works have proposed approaches to learn the parameters of an NCS model based on a learning set. Exact approaches based on mixed integer linear programming ensure that the learning set is best restored, but can only handle datasets of limited size. Heuristic approaches can handle large learning sets, but do not provide any guarantee that the inferred model is the one that best matches the input data. In this paper, we propose an alternative formulation to learn an NCS model. This formulation, based on a SAT problem, guarantees to find a model fully consistent with the learning set (whenever it exists), and is computationally much faster than existing exact MIP approaches.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Multiple Criteria Decision Analysis (MCDA) aims at supporting a decision maker (DM) in making decisions among options described according to various points of view, formally represented by monotone functions called *criteria*. In this paper, decisions are modeled as an *ordinal sorting problem*, where alternatives are to be assigned to a class in the set of predefined ordered classes.

The literature contains several multiple criteria sorting methods which can be distinguished into *(i)* value based sorting methods (see e.g. Greco et al., 2011; Greco et al., 2010; Marichal et al., 2005; Soylu, 2011), *(ii)* outranking based sorting methods (see e.g. Bouyssou and Marchant, 2007a; Bouyssou and Marchant, 2007b; Leroy et al., 2011; Meyer and Olteanu, 2017; Sobrie et al., 2013; Zheng et al., 2014) and, *(iii)* rule based sorting methods (see e.g. Greco et al., 2002; Greco et al., 2016).

We address the problem of ordinal sorting with an outranking based sorting model: the Non-Compensatory Sorting model (NCS, cf.Bouyssou and Marchant, 2007a; Bouyssou and Marchant, 2007b), in which an object is assigned to a class based on its comparison to profiles representing multicriteria frontiers between consecutive classes. NCS assigns an alternative to a category above a profile if it

is at least as good as the profile on a *sufficient* coalition of criteria; the family of sufficient coalitions can be any upset[1] of the set of all subsets of criteria. A particular case of NCS occurs when the family of sufficient coalitions of criteria can be defined using additive criteria weights and a threshold. The literature refers to this additive case as the MR-Sort model (see e.g. Leroy et al., 2011; Sobrie et al., 2013). Both MR-Sort and NCS models are particular cases of the Electre Tri model, a method for sorting alternatives into ordered categories based on an outranking relation (see Roy and Bouyssou, 1993, pp. 389–401 or Bouyssou et al., 2006, pp. 381–385).

Our aim is to learn an NCS model from preference information given in the form of a reference assignment. Such an approach makes it possible to integrate the decision maker preferences into the model without asking her for the preference parameter values. Such indirect elicitation has been developped for Electre Tri (Mousseau and Słowinski, 1998), MR-sort (Leroy et al., 2011), UTADIS (Zopounidis and Doumpos, 2002).

Algorithm 1 describes a general framework that has been widely used (see e.g. Jacquet-Lagrèze and Siskos, 1982; Leroy et al., 2011) in order to leverage the power of generic mathematical programming solvers to learn the parameters of a multicriteria sorting procedure from examples. The workflow is divided into three phases: the problem is *encoded* into a formulation, this formulation

---

* Corresponding author.

E-mail address: vincent.mousseau@centralesupelec.fr (V. Mousseau).

[1] An upset is an upward closed subset of an ordered set, i.e. if *b* is greater than *a* and *a* belongs to an upset, then so does *b*.

Algorithm 1: Learning a model-based classifier.

Input: a tuple of criteria, a tuple of ordered categories, a multicriteria sorting model, an assignment of alternatives to categories

Result: a representation of the assignment in the model, or *None* if the assignment is not representable in the model

1 **encode** the assignment into a formulation Φ
2 try to **solve** the formulation Φ
3   **decode** the solution into a model
4   return the model
5 except NoSolution
6   return *None*

is passed to an external *solver*, and a solution, if found, is then *decoded* into a model. The *faithfulness* of this approach is guaranteed if, and only if:

1. the encoded formulation has a solution as soon as the assignment can be represented in the model;
2. the solver is complete, in the sense that it yields a solution if and only if there is at least one;
3. the decoded model actually represents the assignment.

To the authors' knowledge, until now, general NCS models have been deemed too computationally difficult to address with this approach. Restrictions to MR-Sort have been considered, either in Leroy et al. (2011) with a mixed integer programming (MIP) formulation, but this approach turned out to be inadequate to handle large datasets, or by Sobrie et al. (2013, 2015) using a metaheuristic solving procedure that handles large datasets but offers no guarantee of its completeness (cf. point 2 above). The aim of this paper is to investigate an alternative venue: considering U-NCS, a broad subset of NCS models, that encompasses MR-Sort (for precise definitions of these models, see Section 2.2), and formulating the problem of representing an assignment by a model in U-NCS as a boolean satisfiability problem (SAT). We prove that both the encoding and the decoding satisfy the faithfulness requirements 1 and 3 above. We are thus able to leverage the advances made in the field of generic SAT solvers, to reach unprecedented computational performance in the learning of non-compensatory sorting models.

The paper is organized as follows. In Section 2, we present the notions and concepts related to the formulation of the problem of learning parameters of a non-compensatory sorting model. In Section 3, we develop our binary satisfaction (SAT) problem formulation for inferring an U-NCS model from a learning set, and show it has the desired properties of necessity and sufficiency regarding the representation of an assignment in the U-NCS model. In Section 4, we recall the bases of using a mixed integer formulation to learn the parameters of a MR-Sort model. After that, we propose experiments to assess the pertinence and interest of the SAT formulation in Section 5. In Section 6, we discuss the obtained results. Finally, in Section 7, we conclude by pointing some future interesting perspectives.

2. Position of the problem

In this section, we detail the notions permitting to formulate the problem of learning the parameters of a non-compensatory sorting model. In Section 2.1, we define the vocabulary of ordinal sorting and we formalize the notion of ordinal sorting procedure. We are then able to precisely describe the problem of representing a given assignment in a given ordinal sorting model. In Section 2.2, we proceed by describing the broad class of non-compensatory

sorting models, and two narrower subclasses of particular interest, namely U-NCS and MR-Sort. In Section 2.3, we specify the expected inputs and outputs of the learning problem.

2.1. Vocabulary of multicriteria ordinal sorting

An ordinal sorting problem consists in assigning a category, taken among a given, finite set of *categories* $C^1 \prec \cdots \prec C^p$ ordered by desirability, to *alternatives* described by several attributes.

We assume $\mathcal{N}$ is a finite set of *criteria*, where each criterion $i \in \mathcal{N}$ maps alternatives to values among an ordered set $(\mathbb{X}_i, \leq_i)$, the order relation $\leq_i$ meaning "weakly worse than".[2] An alternative is thus described by a $|\mathcal{N}|$-tuple of multiple criteria values called *profile*. We denote $\mathbb{X} = \prod_{i \in \mathcal{N}} \mathbb{X}_i$ the set of all possible profiles– either describing actual alternatives or virtual ones.

As an analogy with a voting system where criteria would act as voters, subsets of $\mathcal{N}$ are called *coalitions* of criteria. The following function maps a pair of profiles to the coalition of criteria weakly favorable to the former.

$$O_{\mathcal{N}} : \mathbb{X} \times \mathbb{X} \longrightarrow \mathcal{P}(\mathcal{N})$$
$$(x, y) \mapsto \{i \in \mathcal{N} : x_i \geq_i y_i\}$$

When $O_{\mathcal{N}}(x, y) = \mathcal{N}$, the alternative x is at least as good as the alternative y with respect to all criteria, and we say x *weakly dominates y* in the sense of Pareto. Weak dominance defines a partial order Δ on the set of profiles $\mathbb{X}$.

In the remainder of this article, we assume the sets of criteria $\mathcal{N}$, of profiles $\mathbb{X}$ and of categories C are given, and we endeavor to define a *sorting procedure*, a non-decreasing function mapping $\mathbb{X}$ to the set of classes ordered by dominance to $C^1 \prec \cdots \prec C^p$.

2.2. Non-compensatory sorting models

In Bouyssou and Marchant (2007a,b), Bouyssou and Marchant define a set of sorting procedures deemed as *non-compensatory*. We briefly recall the definition of the *non-compensatory sorting (NCS) model*, as well as two specific subsets of this model, *U-NCS* and *MR-Sort*.

All these classes of non-compensatory sorting models, rely on the notions of satisfactory values of the criteria and sufficient coalitions of criteria. They combine into defining the fitness of an alternative: an alternative is deemed fit if it has satisfactory values on a sufficient coalition of criteria.

This notion is straightforward to implement when there are only two categories: the sufficient coalitions $\mathcal{T}$ form an upset of the power set of $\mathcal{N}$ and, for each criterion $i \in \mathcal{N}$, the satisfactory values $\mathcal{A}_i \subset \mathbb{X}_i$ form an upset that can be described by its lower bound $b_i \in \mathbb{X}_i$ – meaning a value is satisfactory if, and only if, it is greater or equal to the threshold b_i, thus defining a *limiting profile* $b \in \mathbb{X}$. With more than two categories, the notions of sufficient coalitions and satisfactory values are declined per category – denoted respectively $\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N}, k \in [1.p-1]}$ and $\langle \mathcal{T}^k \rangle_{k \in [1.p-1]}$. The ordering of the categories $C^1 \prec \cdots \prec C^p$ translates into a nesting of the sufficient coalitions:

$$\forall k \in [1.p-1], \ \mathcal{T}^k \text{ is an upset of } (2^{\mathcal{N}}, \subseteq) \text{ and } \mathcal{T}^1 \supseteq \cdots \supseteq \mathcal{T}^{p-1}$$

(1a)

---

[2] This setting differs from the one described by Bouyssou and Marchant (2007a,b), in the sense that we suppose the attributes describing the alternatives are already sorted by the criteria according to their desirability: here, the order relation on each set $\mathbb{X}_i$ needs not be constructed from holistic preference statements, but is assumed to be established beforehand, e.g. in a previous phase of a decision aiding process structured according to Bouyssou et al. (2006) (this is often the case in applications).

and also a nesting of the satisfactory values:

$\forall i \in \mathcal{N}, \ \forall k \in [1..p-1], \ \mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \leq_i)$

and $\mathcal{A}_i^1 \supseteq \cdots \supseteq \mathcal{A}_i^{p-1}$ (1b)

Condition (1b) translates into an ordering of the values $\langle b_i^k \rangle_{k \in [1..p-1]}$ for a given criterion $i \in \mathcal{N}$, or an ordering of the limiting profiles:

$b^1, \ldots, b^{p-1}$ is a non-decreasing sequence of $(\mathbb{X}, \Delta)$ (1c)

For convenience, these sequences are augmented with trivial elements on both ends: $\mathcal{T}^0 := \mathcal{P}(\mathcal{N})$, $T^p := \emptyset$, $\forall i \in \mathcal{N} \mathcal{A}_i^0 = \mathbb{X}_i$, $\mathcal{A}_i^p = \emptyset$, $b^0 := \bot$, $b^p := \top$.

Definition 1 (Non-compensatory sorting NCS, Bouyssou and Marchant, 2007b). Given a set of criteria $\mathcal{N}$ and an ordered set of categories $C^1 \prec \cdots \prec C^p$, for all pairs of tuples $(\langle b \rangle, \langle \mathcal{T} \rangle)$ where $\langle b \rangle$ satisfies (1c) and $\langle \mathcal{T} \rangle$ satisfies (1a), the sorting function $NCS_{(\langle b \rangle, \langle \mathcal{T} \rangle)}$ maps a profile $x \in \mathbb{X}$ to the category C^k such that $O_{\mathcal{N}}(x, b^k) \in \mathcal{T}^k$ and $O_{\mathcal{N}}(x, b^{k+1}) \notin \mathcal{T}^{k+1}$.

The set of preference parameters – all the pairs $(\langle b \rangle, \langle \mathcal{T} \rangle)$ satisfying (1a) and (1c) – can be considered too wide and too unwieldy for practical use in the context of a decision aiding process. Therefore, following (Bouyssou and Marchant, 2007b), one may consider to restrict either the sequence of limiting profiles, or the sequence of sufficient coalitions. In order to remain compatible with Electre Tri, we elect the latter.

Definition 2 (Non-compensatory sorting with a unique set of sufficient coalitions U-NCS). Given a set of criteria $\mathcal{N}$ and an ordered set of categories $C^1 \prec \cdots \prec C^p$, for all pairs $(\langle b \rangle, \mathcal{T})$ where the tuple $\langle b \rangle$ satisfies (1c) and $\mathcal{T}$ is an upset of coalitions, the sorting function $U\text{-}NCS_{(\langle b \rangle, \mathcal{T})}$ maps a profile $x \in \mathbb{X}$ to the category C^k such that $O_{\mathcal{N}}(x, b^k) \in \mathcal{T}$ and $O_{\mathcal{N}}(x, b^{k+1}) \notin \mathcal{T}$.

A further restriction of U-NCS is of particular interest: in the MR-Sort model, introduced in Leroy et al. (2011), the sufficient coalitions are represented in a compact form which is more amenable to linear programming. As an analogy to a voting setting, each criterion $i \in \mathcal{N}$ may be assigned a *voting power* $w_i \geq 0$ so that a given coalition of criteria $B \subseteq \mathcal{N}$ is deemed sufficient if, and only if, its combined voting power $\Sigma_{i \in B} w_i$ is greater than a given *qualification threshold* λ.

Definition 3 (majority rule sorting MR-Sort). Given a set of criteria $\mathcal{N}$, the majority rule MR maps a pair $(\langle w \rangle, \lambda)$, where $\langle w \rangle$ is a $|\mathcal{N}|$-tuple of nonnegative real numbers and λ a nonnegative real number, to an upset $MR(\langle w \rangle, \lambda)$ of the power set of $\mathcal{N}$ defined by the relation:

$\forall B \subseteq \mathcal{N}, \ B \in MR(\langle w \rangle, \lambda) \iff \sum_{i \in B} w_i \geq \lambda$ (MR)

Given, in addition, a set of categories $C^1 \prec \cdots \prec C^p$, for all triples $(\langle b \rangle, \langle w \rangle, \lambda)$ where the tuple $\langle b \rangle$ satisfies (1c), $\langle w \rangle$ is a $|\mathcal{N}|$-tuple of nonnegative real numbers and λ a nonnegative real number, $MR\text{-}Sort_{(\langle b \rangle, \langle w \rangle, \lambda)}$ is the sorting function $U\text{-}NCS_{(\langle b \rangle, MR(\langle w \rangle, \lambda))}$.

Example 1. Terry is a journalist and prepares a car review for a special issue. He considers a number of popular car models, and wants to sort them in order to present a sample of cars "selected for you by the editorial board" to the readers.

This selection is based on 4 criteria : cost (€), acceleration (measured by the time, in seconds, to reach 100 km/h from full stop – lower is better), braking power and road holding, both measured on a qualitative scale ranging from 1 (lowest performance) to 4 (best performance). The performances of six models are described in Table 1:

In order to assign these models to a class among $C^{1\star}$ (average) $\prec C^{2\star}$ (good) $\prec C^{3\star}$ (excellent), Terry considers a NCS model:

Table 1
Performance table.

Model	Cost	Acceleration	Braking	Road holding
m_1	16 973	29	2.66	2.5
m_2	18 342	30.7	2.33	3
m_3	15 335	30.2	2	2.5
m_4	18 971	28	2.33	2
m_5	17 537	28.3	2.33	2.75
m_6	15 131	29.7	1.66	1.75

Table 2
Categorization of performances.

Model	Cost	Acceleration	Braking	Road holding
m_1	**	**	***	**
m_2	*	*	**	***
m_3	***	*	*	**
m_4	*	***	**	**
m_5	*	***	**	***
m_6	***	**	*	*

Table 3
Model Assignments.

Model	Assignment
m_1	**
m_2	*
m_3	**
m_4	**
m_5	***
m_6	*

- Where the values on each criterion are sorted between $1^\star$ (average) and $2^\star$ (good) by the following profiles: $b^{1\star}_{cost} = 17\,250$, $b^{1\star}_{acceleration} = 30$, $b^{1\star}_{braking} = 2.2$, $b^{1\star}_{road\ holding} = 1.9$. The boundary between $2^\star$ and $3^\star$ (excellent) is fixed by the profiles: $b^{2\star}_{cost} = 15\,500$, $b^{2\star}_{acceleration} = 28.8$, $b^{2\star}_{braking} = 2.5$, $b^{2\star}_{road\ holding} = 2.6$. Fig. 1 and Table 2 depict the performance of the six alternatives.

- These appreciations are then aggregated by the following rule: *an alternative is categorized good or excellent if it is good or excellent on cost or acceleration, and good or excellent on braking or road holding. It is categorized excellent if it is excellent on cost or acceleration, and excellent on braking or road holding.* Being excellent on some criterion does not really help to be considered good overall, as expected from a non-compensatory model. Sufficient coalitions are represented on Fig. 2.

Finally, the model yields the following assignments (Table 3):

2.3. The disaggregation paradigm: Learning preference parameters from assignment examples

For a given decision situation, assuming the NCS model is relevant to structure the DM's preferences, what parameters should be selected to fully specify the NCS model that corresponds to the DM viewpoint? An option would be to simply ask the decision maker to describe, to her best knowledge, the limit profiles between classes and to enumerate the minimal sufficient coalitions. In order to get this information as quickly and reliably as possible, an analyst could make good use of the *model-based elicitation strategy* described in Belahcène et al. (2017b), as it permits to obtain these parameters by asking the decision maker to only provide holistic preference judgment – should some (fictitious) alternative be assigned to some category – and builds the shortest questionnaire.

We opt for a more indirect setup, close to a machine learning paradigm, where a set of reference assignment is given and as-
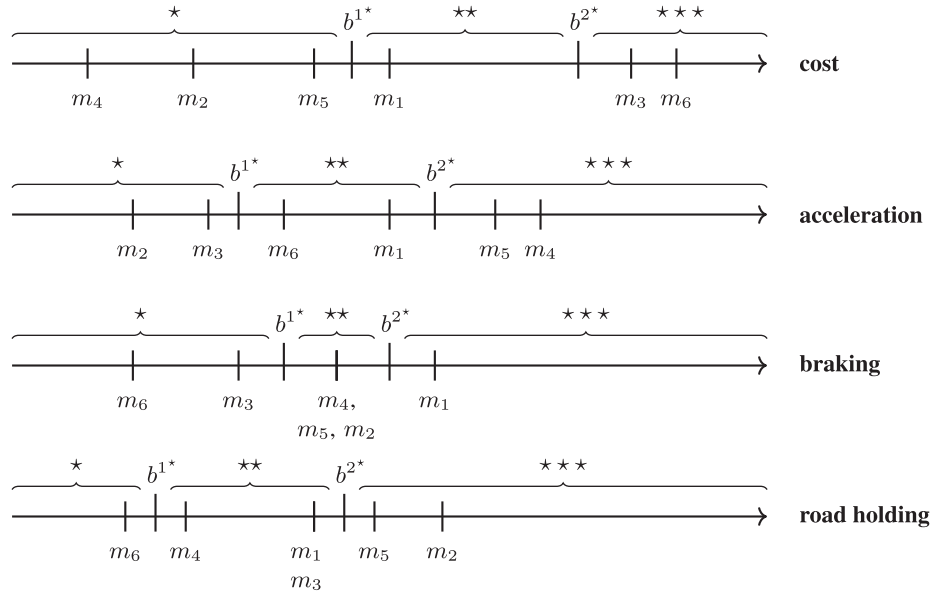
Fig. 1. Representation of performances w.r.t. category limits.
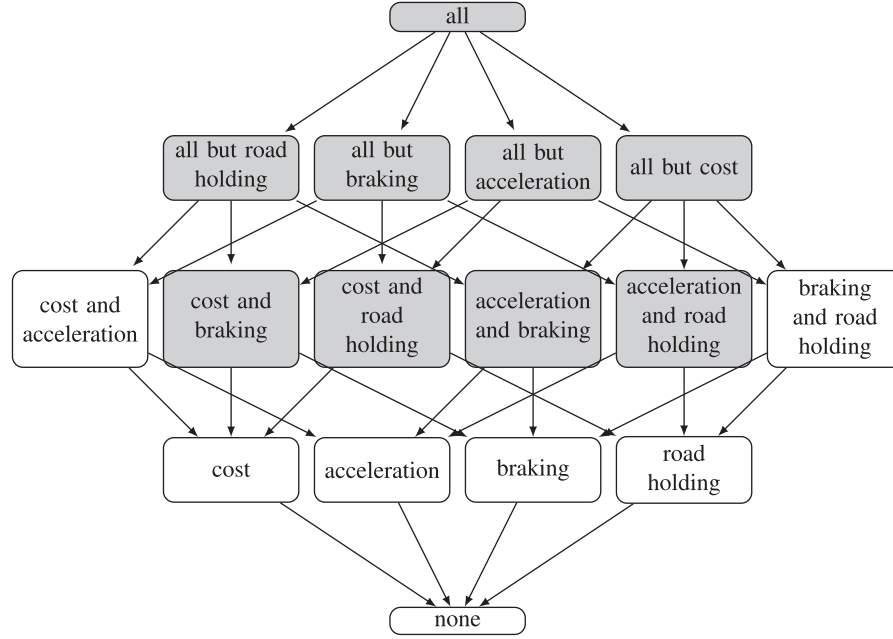
Fig. 2. Sufficient (grey) and insufficient (white) coalitions of criteria. Arrows denote strength - pointing towards the weaker.

sumed to describe the decision maker's point of view, and the aim is to *extend* these assignments with a NCS model. In this context, we usually refer to an *assignment* as a function mapping a subset of *reference alternatives* $\mathbb{X}^\star \subset \mathbb{X}$ to the ordered set of classes $C^1 \prec \cdots \prec C^p$. These reference alternatives highlight values of interest on each criterion $i \in \mathcal{N}$, $\mathbb{X}_i^\star := \bigcup_{x \in \mathbb{X}^\star} \{x_i\}$. We are looking for suitable preference parameters specifying a non-compensatory sorting model, i.e. a tuple of profiles $\langle b \rangle$ satisfying (1c) and an up-set of coalitions $\mathcal{T} \subset 2^{\mathcal{N}}$ (respectively, non-negative voting parameters $(\langle w \rangle, \lambda)$ of a majority rule) so that U-NCS$_{(\langle b \rangle, \mathcal{T})}$ (respectively, MR-Sort$_{(\langle b \rangle, \langle w \rangle, \lambda)}$) maps each reference alternative $x \in \mathbb{X}^\star$ to its assigned class $A(x)$.

Throughout this paper, we assume the expression of preference is free of noise. We are only interested in determining if the given assignment can be represented in the non-compensatory sorting model.

3. SAT formulation for learning NCS

In this section, we begin by giving a brief reminder of some key concepts regarding boolean satisfiability problems (SAT). Then, we proceed by describing the pivotal contribution of this work: the encoding of the problem of representing a given assignment in the U-NCS model as a SAT problem. We conclude this section by providing the decoding procedure that prove this SAT formulation is equivalent to the original problem, and can be used in the context of Algorithm 1 together with a SAT solver.

3.1. Boolean satisfiability (SAT)

A boolean satisfaction problem consists in a set of boolean variables V and a logical proposition about these variables f: $\{0, 1\}^V \to \{0, 1\}$. A solution $v^\star$ is an assignment of the vari-

ables mapped to 1 by the proposition: $f(v^\star) = 1$. A binary satisfaction problem for which there exists at least one solution is *satisfiable*, else it is *unsatisfiable*. Without loss of generality, the proposition f can be assumed to be written in conjunctive normal form: $f = \bigwedge_{c \in \mathcal{C}} c$, where each *clause* $c \in \mathcal{C}$ is itself a disjunction in the variables or their negation $\forall c \in \mathcal{C}, \exists c^+, c^- \in \mathcal{P}(V) : c = \bigvee_{v \in c^+} v \vee \bigvee_{v \in c^-} \neg v$, so that a solution satisfies at least one condition (either positive or negative) of every clause.

The models presented hereafter make extensive use of clauses where there is only one non-negated variable (a subset of *Horn clauses*): $a \vee \neg b_1 \vee \cdots \vee \neg b_n$, which represent the logical implication $(b_1 \wedge \cdots \wedge b_n) \Rightarrow a$.

It is known since Cook's theorem (Cook, 1971) that the Boolean satisfiability problem is NP-complete. Consequently, unless $P = NP$, we should not expect to solve generic SAT instances quicker than exponential time in the worst case. Nevertheless, efficient and scalable algorithms for SAT have been – and still are – developed, and are sometimes able to handle problem instances involving tens of thousands of variables and millions of clauses in a few seconds (see e.g. Biere et al., 2009; Moskewicz et al., 2001).

3.2. A SAT encoding of a given assignment in U-NCS

Definition 4 (SAT encoding for U-NCS). Let $A : \mathbb{X}^\star \to C^1 \prec \cdots \prec C^p$ an assignment. We define the boolean function ϕ_A^{SAT} with variables:

- $x_{i, h, k}$, indexed by a criterion $i \in \mathcal{N}$, a frontier between classes $1 \le h \le p - 1$, and a value $k \in \mathbb{X}_i^\star$ taken on criterion i by a reference alternative,
- y_B indexed by a coalition of criteria $B \subseteq \mathcal{N}$

as the conjunction of clauses:

(2a) For all criteria $i \in \mathcal{N}$, for all frontiers between adjacent classes $1 \le h \le p - 1$, for all ordered pairs of values $k < k' \in \mathbb{X}_i^\star$:

$$x_{i,h,k'} \vee \neg x_{i,h,k} \tag{2a}$$

(2b) For all criteria $i \in \mathcal{N}$, for all ordered pairs of frontiers $1 \le h < h' \le p - 1$, for all values $k \in \mathbb{X}_i^\star$:

$$x_{i,h,k} \vee \neg x_{i,h',k} \tag{2b}$$

(2c) For all ordered pairs of coalitions $B \subset B' \subseteq \mathcal{N}$:

$$y_{B'} \vee \neg y_B \tag{2c}$$

(2d) For all coalitions $B \subseteq \mathcal{N}$, for all frontiers $1 \le h \le p - 1$, for all $u \in \mathbb{X}^\star : A(u) = C^{h-1}$ (i.e. reference alternatives just below the frontier) :

$$\left(\bigvee_{i \in B} \neg x_{i,h,u_i} \right) \vee \neg y_B \tag{2d}$$

(2e) For all coalitions $B \subseteq \mathcal{N}$, for all frontiers $1 \le h \le p - 1$, for all $a \in \mathbb{X}^\star : A(a) = C^h$ (i.e. reference alternatives just above the frontier) :

$$\left(\bigvee_{i \in B} x_{i,h,a_i} \right) \vee y_{\mathcal{N} \setminus B} \tag{2e}$$

Clauses of types (2a), (2b) and (2c) are easily interpreted as enforcers of some monotonicity conditions inherent to ordinal sorting and to the parameters of the U-NCS model:

(2a) *Ascending scales* – if $k < k' \in \mathbb{X}_i^\star$ and k is above some threshold b_i^h, then so is k'. It is necessary and sufficient to consider the clauses where k and k' are consecutive values of $\mathbb{X}_i^\star$.

(2b) *Hierarchy of profiles* – if $1 \le h < h' \le p - 1$ and $k \in \mathbb{X}_i^\star$ is above the threshold $b_i^{h'}$, then it is also above b_i^h. It is necessary and sufficient to consider the clauses where $h' = h + 1$.

(2c) *Coalitions strength* – if a coalition $B \subseteq \mathcal{N}$ is sufficient, then any coalition $B' \supseteq B$ containing B is also sufficient. It is necessary and sufficient to consider the clauses where the coalition B' contains exactly one more criterion than B, corresponding to the edges represented on Fig. 1.

Clauses of types (2d) and (2e) ensure the correct representation of all reference alternatives contained by the assignment A in the U-NCS model. They rely on the following lemmas.

Lemma 1. *Let $A : \mathbb{X}^\star \to C^1 \prec \cdots \prec C^p$ an assignment extended by a U-NCS model with profiles $\langle b \rangle$ and sufficient coalitions $\mathcal{T}$. If $B \subseteq \mathcal{N}$ is a coalition of criteria such that, there is an alternative $x \in \mathbb{X}^\star$ stronger than the upper frontier of its class $b^{A(x)+1}$ on every criterion in B, then this coalition is not sufficient.*

$$\forall B \subseteq \mathcal{N}, \ [\exists x \in \mathbb{X}^\star \ : \ \forall i \in B, \ x_i \ge b_i^{A(x)+1}] \Rightarrow B \notin \mathcal{T}$$

Proof. Let A an assignment, $(\langle b \rangle, \mathcal{T})$ correct U-NCS parameters , B a coalition of criteria and x an alternative that satisfy the premises, and suppose B is sufficient. The alternative x would be better than the boundary $b^{A(x)+1}$ and so would be assigned to a class strictly better than $A(x)$, and the NCS model with parameters b and $\mathcal{T}$ would not extend the assignment. $\square$

Clauses of type (2d) leverage Lemma 1 to ensure *alternatives are outranked by the boundary above them*, relating insufficient coalitions to the strong points of an alternative.

Lemma 2. *Let $A : \mathbb{X}^\star \to C^1 \prec \cdots \prec C^p$ an assignment extended by a U-NCS model with profiles $\langle b \rangle$ and sufficient coalitions $\mathcal{T}$. If $B \subseteq \mathcal{N}$ is a coalition of criteria such that, there is an alternative $x \in \mathbb{X}^\star$ weaker than the lower frontier of its class $b^{A(x)}$ on every criterion in B, then the complementary coalition is sufficient.*

$$\forall B \subseteq \mathcal{N}, \ [\exists x \in \mathbb{X}^\star \ : \ \forall i \in B, \ x_i < b_i^{A(x)}] \Rightarrow (\mathcal{N} \setminus B) \in \mathcal{T}$$

Proof. Let A an assignment, $(\langle b \rangle, \mathcal{T})$ correct U-NCS parameters, B a coalition of criteria and x an alternative that satisfy the premises, and suppose $\mathcal{N} \setminus B$ is insufficient. The set of criteria on which the alternative x would be better than the boundary $b^{A(x)}$ is a subset of $\mathcal{N} \setminus B$, and would thus be considered insufficient. Hence, x would be assigned to a class strictly worse than $A(x)$, and the NCS model with parameters b and $\mathcal{T}$ would not extend the assignment. $\square$

Clauses of type (2e) leverage Lemma 2 to ensure *alternatives outrank the boundary below them*, relating the weak points of an alternative to a complementary insufficient coalition.

We are now able to describe the decoding function required by Algorithm 1 and prove the faithfulness of both the encoding and the decoding.

3.3. Faithfulness of the SAT representation

Theorem 1 (from a U-NCS model representing an assignment to a solution of the SAT formulation). *Given an assignment $A : \mathbb{X}^\star \to C^1 \prec \cdots \prec C^p$, if the tuple of profiles $\langle b \rangle$ satisfies (1c), the set $\mathcal{T}$ is an upset of coalitions of criteria, and the sorting function U-NCS$_{\langle b \rangle, \mathcal{T}}$ extends A, then the binary tuple:*

- $x_{i, h, k}$, *indexed by a criterion $i \in \mathcal{N}$, a frontier between classes $1 \le h \le p - 1$, and a value $k \in \mathbb{X}_i^\star$ taken on criterion i by some reference alternative, and defined by* $x_{i,h,k} = \begin{cases} 1, & \text{if } k \ge b_i^h \\ 0, & \text{else} \end{cases}$
- y_B *indexed by a coalition of criteria $B \subseteq \mathcal{N}$ and defined by* $y_B = \begin{cases} 1, & \text{if } B \in \mathcal{T} \\ 0, & \text{else} \end{cases}$

is mapped to 1 by the Boolean function ϕ_A^{SAT}.

Proof. The clauses (2a) are satisfied because if $k < k'$ and k is above some threshold b^h, then so is k'. The clauses (2b) are satisfied because the frontier profiles $\langle b \rangle$ are assumed to satisfy (1c) (hence, if a given value is above some threshold $b_i^{h'}$, then it is also above inferior thresholds b_i^h for $h < h'$). The clauses (2c) are satisfied because $\mathcal{T}$ is assumed to be an upset (hence, if a coalition is deemed sufficient, then so are wider coalitions). If the NCS model with profiles b^h and sufficient coalitions $\mathcal{T}$ extends the given assignments, then clauses (2d) are satisfied – else, by Lemma 1, one of the alternative $u \in \mathbb{X}^\star$ assigned to the class C^{h-1} would outrank the profile b^h on a sufficient coalition of criteria. So are clauses (2e) - else, by Lemma 2, one alternative $a \in \mathbb{X}^\star$ assigned to class C^h would not outrank the profile b^h, as the set of criteria on which a is better than b^h would be smaller than some insufficient coalition. □

Corollary 1 (Faithful encoding). *Let A be an assignment $A : \mathbb{X}^\star \to C^1 \prec \cdots \prec C^p$. If ϕ_A^{SAT} is unsatisfiable, then A cannot be represented in the model U-NCS.*

Theorem 2 (Decoding a solution of the SAT formulation into a U-NCS model). *Given an assignment $A : \mathbb{X}^\star \to C^1 \prec \cdots \prec C^p$, if the binary tuple:*

- *$x_{i, h, k}$, indexed by a criterion $i \in \mathcal{N}$, a frontier between classes $1 \le h \le p - 1$, and a value $k \in \mathbb{X}_i^\star$ taken on criterion i by a reference alternative,*
- *y_B indexed by a coalition of criteria $B \subseteq \mathcal{N}$*

satisfies $\phi_A^{SAT}(x, y) = 1$, then the profiles $\langle b \rangle$ defined by $b_i^h := \min\{k \in \mathbb{X}_i^\star : x_{i,h,k} = 1\}$ satisfy (1c), the set of coalitions $\mathcal{T} := \{B \subseteq \mathcal{N} : y_B = 1\}$ is an upset, and the sorting function U-NCS$_{(\langle b \rangle, \mathcal{T})}$, extends the assignment A.

Proof. Clauses (2a) ensure that $k' \ge k \Rightarrow x_{i,h,k'} \ge x_{i,h,k}$, so that $x_{i,h,k} = 1 \iff k \ge b_i^h$. Clauses (2b) ensure the tuple of profiles $\langle b \rangle$ satisfies (1c). Clauses (2c) ensure the set $\mathcal{T}$ is an upset of coalitions. The sorting function U-NCS$_{(\langle b \rangle, \mathcal{T})}$ extends the given assignment because, for each reference alternative $s \in \mathbb{X}^\star$, there is a clause (2e) that ensures s outranks the lower frontier of its class (if $A(s) \succ C^1$), and a clause (2d) that ensures s does not outrank the upper frontier of its class (if $A(s) \prec C^p$). □

Corollary 2 (Faithfullness of the SAT representation). *The assignment A can be represented in the model U-NCS if, and only if, ϕ_A^{SAT} is satisfiable.*

4. Learning MR-sort using mixed integer programming

Learning the parameters of an MR-Sort model using mixed integer programming has been studied in Leroy et al. (2011). We recall here the method used in Leroy et al. (2011) in order to obtain the mixed integer program (MIP) formulation that infers an MR-Sort model on the basis of examples of assignments.

With MR-Sort (see Definition 3), the condition for an alternative $a \in \mathbb{X}^\star$ to be assigned to a category C^h reads:

$$a \in C^h \iff \begin{cases} \sum_{i=1}^n c_{a,i}^{h-1} & \ge \lambda \\ \sum_{i=1}^n c_{a,i}^h & < \lambda \end{cases} \text{ with } c_{a,i}^k = \begin{cases} w_i & \text{if } a_i \ge b_i^k, \\ 0 & \text{otherwise.} \end{cases}$$

The linearization of these constraints induces the use of binary variables. For each variable $c_{a,i}^k$, with $k = \{h - 1, h\}$, we introduce a binary variable $\delta_{a,i}^k$ that is equal to 1 when the performance of $a \in \mathbb{X}^\star$ is at least as good as or better than the performance of b^k on the criterion i and 0 otherwise. For an alternative a assigned to a category C^h with $2 \le h \le p - 1$, it introduces $2n$ binary variables. For alternatives assigned to one of the extreme categories,

the number of binary variables is divided by two. The value of each variable $\delta_{a,i}^k$ is obtained thanks to the following constraints:

$$M(\delta_{a,i}^k - 1) \le a_i - b_i^k < M \cdot \delta_{a,i}^k \tag{3a}$$

in which M is an arbitrary large positive constant[3]. The value of $c_{a,i}^k$ are finally obtained thanks to the following constraints:

$$\begin{cases} 0 \le & c_{a,i}^k & \le w_i, \\ \delta_{a,i}^k - 1 + w_i \le & c_{a,i}^k & \le \delta_{a,i}^k. \end{cases} \tag{3b}$$

The dominance structure on the set of profiles is ensured by the following constraints:

$$\forall i \in \mathcal{N}, h = \{2, \ldots, p - 1\}, \ b_i^h \ge b_i^{h-1} \tag{3c}$$

As the Eq. (MR) defining the majority rule is homogenous, the coefficients $\langle w \rangle$ and λ can be multiplied by any positive constant without modifying the upset of coalitions they represent. Thus, the following normalization constraint can be added without loss of generality:

$$\sum_{i=1}^n w_i = 1. \tag{3d}$$

To obtain a MIP formulation, the next step consists to define an objective function. In Leroy et al. (2011), two objective functions are considered, one of which consists in maximizing the robustness of the assignments. It is done by adding continuous variables $x_a, y_a \in \mathbb{R}$ for each alternative $a \in \mathbb{X}^\star$ such that:

$$\begin{cases} \sum_{i=1}^n c_{a,i}^{h-1} & = \lambda + x_a, \\ \sum_{i=1}^n c_{a,i}^h & = \lambda - y_a. \end{cases} \tag{3e}$$

The objective function consists in optimizing a slack variable α that is constrained by the values of the variables x_a and y_a as follows:

$$\forall a \in \mathbb{X}^\star, \begin{cases} \alpha & \le x_a, \\ \alpha & \le y_a. \end{cases} \tag{3f}$$

The combination of the objective function and all the constraints listed above leads to MIPs that can be found in Leroy et al. (2011).

Definition 5 (MIP-O formulation for MR-Sort). Given an assignment A, we denote ϕ_A^{MIP-O} the mixed linear program with decision variables α, λ, $\langle b_i^h \rangle_{i \in \mathcal{N}, k \in [1,p-1]}$, $\langle w_i \rangle_{i \in \mathcal{N}}$, $\langle c_{a,i}^h \rangle_{i \in \mathcal{N}, a \in \mathbb{X}^\star, h \in \{A(a)-1, A(a)\}}$, $\langle x_a \rangle_{a \in \mathbb{X}^\star}$, $\langle y_a \rangle_{a \in \mathbb{X}^\star} \in \mathbb{R}^+$ and $\langle \delta_{a,i}^h \rangle_{i \in \mathcal{N}, a \in \mathbb{X}^\star, h \in \{A(a)-1, A(a)\}} \in \{0, 1\}$, consisting in minimizing the objective α, subject to the constraints (3a)–(3f).

Theorem 3 (Faithfulness of the MIP-O formulation Leroy et al., 2011). *An assignment A can be represented in the model MR-Sort if, and only if, ϕ_A^{MIP-O} is feasible. If the tuple $\langle \alpha, \lambda, b, w, c, x, y, \delta \rangle$ is a feasible solution of ϕ_A^{MIP-O}, then the tuple of profiles b, the tuple of voting powers w and the majority threshold λ are suitable parameters of a MR-Sort model that extends the assignment A.*

We are looking to compare this state-of-the-art formulation to the boolean satisfiability formulation we propose in the next section in terms of computational efficiency, and in terms of quality of the result. Yet, we suspect the two approaches differ in too many aspects to be meaningfully compared. The MIP-O formulation is based on a numerical representation of the problem, considers the set of every MR-Sort model extending the assignment, and selects the best according to the objective function – here, returning the model that gives the sharpest difference in voting weights between sufficient and insufficient coalitions of criteria. Meanwhile, the SAT formulation is based on a logical representation of the problem,

---

[3] $M > Max_{i \in \mathcal{N}} max \mathbb{X}_i$

considers the wider set of every U-NCS model extending the assignment, and randomly yields a suitable model. In order to be able to credit the effects we would observe to the correct causes, we introduce a third formulation, called MIP-D, that helps bridging the gap between MIP-O and SAT. MIP-D is formally a mixed integer program with a null objective function. This trick enables us to use the optimization shell of the MIP formulations to express a decision problem assessing the satisfiability of the constraints, and yielding a random solution (which, in our context, represents a particular MR-Sort model), rather than looking for the best one in the sense of the objective function. Another instance of this configuration, where an optimization problem is compared to its feasibility version, can be found in Dickerson et al. (2014). Here, it should be noted that the MIP-D formulation is not exactly the feasibility version of MIP-O, as insufficient coalitions of criteria are characterized by a strict comparison. The optimization version circumvents this obstacle by maximizing the contrast in normalized voting power between sufficient and insufficient coalitions, while the feasibility version addresses it by leaving the total weight unconstrained, but requiring the minimal difference between sufficient and insufficient coalitions is at least one vote. This slight difference might account for some divergence of behavior we observe during our experiment (see Section 5, and particularly 5.3).

Definition 6 (MIP-D formulation for MR-Sort). We denote MIP-D the mixed linear program with decision variables $\langle b_i^k \rangle_{i \in \mathcal{N}, k \in [1..p-1]}$, $\langle w_i \rangle_{i \in \mathcal{N}}$, λ, $\langle x_a \rangle_{a \in \mathbb{X}^\star}$, $\langle y_a \rangle_{a \in \mathbb{X}^\star}$, $\langle c_{a,i}^h \rangle_{i \in \mathcal{N}, a \in \mathbb{X}^\star, h \in \{A(a)-1, A(a)\}} \in \mathbb{R}^+$ and $\langle \delta_{a,i}^h \rangle_{i \in \mathcal{N}, a \in \mathbb{X}^\star, h \in \{A(a)-1, A(a)\}} \in \{0, 1\}$, consisting in minimizing the objective 0, subject to the constraints (3a), (3b), (3c), (3e) and (3g), where:

$$\forall a \in \mathbb{X}^\star, \begin{cases} 1 & \leq x_i, \\ 1 & \leq y_i. \end{cases} \tag{3g}$$

Theorem 4 (Faithfulness of the MIP-D formulation). *An assignment A can be represented in the model MR-Sort if, and only if, ϕ_A^{MIP-D} is feasible. If the tuple $\langle \lambda, b, w, c, x, y, \delta \rangle$ is a feasible solution of ϕ_A^{MIP-D}, then the tuple of profiles b, the tuple of voting powers w and the majority threshold λ are suitable parameters of a MR-Sort model that extends the assignment A.*

Proof. This theorem results from Theorem 3, with only minor changes to the constraints. As noted previously, the normalization constraint (3d) has no effect on the feasibility of the problem. Instead, constraints (3g) ensure we are looking for voting parameters large enough to have at least a difference of one unit between the votes gathered by any sufficient coalition on the one hand and any insufficient coalition on the other hand. □

5. Implementation

In this section, we study the performance of the formulation proposed in Section 3, both intrinsic and comparative with respect to state-of-the-art techniques. We implement Algorithm 1, using a state-of-the-art SAT solver, in order to solve instances of the problem of learning a U-NCS model, given the assignment of a set of reference alternatives. We also implement two formulations relying on Mixed Integer Programming, presented in Section 4, using an adequate solver. We begin by describing our experimental protocol, with some implementation details. Then, we provide the results of the experimental study concerning the computation time of our program, and particularly the influence the size of the learning set, the number of criteria, and the number of classes, as well as elements of comparison between the three approaches.

5.1. Experimental protocol and implementation details

The algorithm we test takes as an input the assignment of a set of alternatives $\mathbb{X}^\star$, each described by a performance tuple on a set of criteria $\mathcal{N}$, to a set of classes $C^1 \prec \cdots \prec C^p$.

The performance of the solvers needs to be measured in practice, by solving actual instances of the problem and reporting the computation time required. This experimental study is run on an ordinary laptop with Windows 7 (64 bit) equipped with an Intel Core i7-4600 CPU at 2.1 GHz and 8 GB of RAM.

5.1.1. Dataset generation

In the scope of this paper, we only consider to use a carefully crafted, random dataset as an input. On the one hand, the algorithm we describe is not yet equipped with the capability to deal with noisy inputs, so we do not consider feeding it with actual preference data, such as the one found in preference learning benchmarks (Fürnkranz and Hüllermeier, 2011). On the other hand, using totally random, unstructured, inputs makes no sense in the context of algorithmic decision. In order to ensure the preference data we are using makes sense, we use a decision model to generate it, and, in particular, a model compatible with the noncompensatory stance we are postulating. Precisely, we use a MR-Sort model for generating the learning set, a model that particularizes NCS and U-NCS by postulating the set of sufficient coalitions possess an additive structure (see Section 2.2). This choice ensures the three formulations we are using should succeed in finding the parameters of a model extending the reference assignment.

When generating a dataset, we consider the number of criteria $|\mathcal{N}|$, the number of classes p, and the number of reference alternatives $|\mathbb{X}^\star|$ as parameters. We consider all criteria take continuous values in the interval [0,1], which is computationally more demanding for our algorithm than the case where one criterion has a finite set of values. We generate a set of ascending profiles $\langle b \rangle$ by uniformly sampling $p-1$ numbers in the interval [0,1] and sorting them in ascending order, for all criteria. We generate voting weights $\langle w \rangle$ by sampling $|\mathcal{N}|-1$ numbers in the interval [0,1], sorting them, and using them as the cumulative sum of weights. λ is then randomly chosen with uniform probability in the interval]0.5,1[. Finally, we sample uniformly $|\mathbb{X}^\star|$ tuples in $[0, 1]^\mathcal{N}$, defining the performance table of the reference alternative, and assign them to classes in $C^1 \prec \cdots \prec C^p$ according to the model $\mathcal{M}^0 :=$MR-Sort$_{\langle b \rangle, \langle w \rangle, \lambda}$ with the generated profiles, voting weights, and qualified majority threshold.

5.1.2. Solving the SAT problem

We then proceed accordingly to Algorithm 1, translating the assignment into a binary satisfaction problem, described by sets of variables and clauses, as described by Definition 4. This binary satisfaction problem is written in a file, in DIMACS format[4], and passed to a command line SAT solver - CryptoMiniSat 5.0.1 (Soos, 2016), winner of the incremental track at SAT Competition 2016[5], released under the MIT license. If the solver finds a solution, then it is converted into parameters $(\langle b^{SAT} \rangle, \mathcal{T}^{SAT})$ for a U-NCS model, as described by Theorem 2. The model $\mathcal{M}^{SAT} :=$U-NCS$_{\langle b^{SAT} \rangle, \mathcal{T}^{SAT}}$ yielded by the program is then validated against the input. As the ground truth $\mathcal{M}^0$ used to seed the assignment is, by construction, a MR-Sort model and therefore a U-NCS model, Theorem 1 applies and we expect the solver to always find a solution. Moreover, as Theorem 2 applies to the solution yielded, we expect the U-NCS model returned by the program should always succeed at extending the assignment provided.
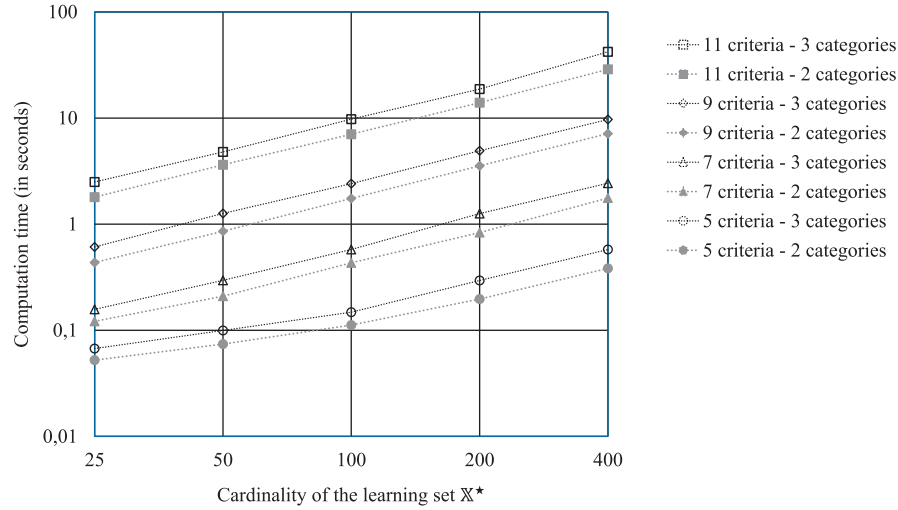
---

[4] http://www.satcompetition.org/2009/format-benchmarks2009.html .

[5] http://baldur.iti.kit.edu/sat-competition-2016/ .

Fig. 3. Computation time by size of the learning set.

5.1.3. Solving the MIP problems.

We transcribe the problem consisting of finding a MR-Sort model extending the assignment with parameters providing a good contrast into a mixed integer linear optimization problem described extensively in Section 4 that we refer to as *MIP-O*, where O stands for *optimization*. In order to bridge the gap between this optimization stance and the boolean satisfiability approach that is only preoccupied with returning any model that extends the given assignment, we also transcribe the problem consisting of finding *some* MR-Sort model extending the assignment into a MIP feasibility problem (optimizing the null function over an adequate set of constraints), also described in Section 4 that we refer to as *MIP-D*, where D stands for *decision*. These MIP problems are solved with Gurobi 7.02, with factory parameters except for the cap placed on the number of CPU cores devoted to the computation (two), in order to match a similar limitation with the chosen version of the SAT solver. When the solver succeeds in finding a solution before the time limit – set to one hour – the sorting models returned are called $\mathcal{M}^{\mathrm{MIP-O}}$ and $\mathcal{M}^{\mathrm{MIP-D}}$, respectively.

5.1.4. Evaluating the ability of the inferred models to restore the original one.

In order to appreciate how "close" a computed model $\mathcal{M}^c \in \{\mathcal{M}^{\mathrm{SAT}}, \mathcal{M}^{\mathrm{MIP-D}}, \mathcal{M}^{\mathrm{MIP-O}}\}$ is to the ground truth $\mathcal{M}^0$ from which the assignment examples were generated, we proceed as follows: we sample a large set of n profiles in $\mathbb{X} = [0,1]^{\mathcal{N}}$ and compute the assignment of these profiles according to the original and computed MR-Sort models ($\mathcal{M}^0$ and $\mathcal{M}^c$). On this basis, we compute *err – rate* the proportion of "errors", *i.e.* tuples which are not assigned to the same category by both models.

5.2. Intrinsic performance of the SAT formulation

We run the experimental protocol described above by varying the various values of the parameters governing the input. In order to assess the intrinsic performance of Algorithm 1 we consider all the combinations where

- the number of criteria $|\mathcal{N}|$ is chosen among {5, 7, 9, 11};
- the number of reference alternatives $|\mathbb{X}^\star|$ is chosen among {25, 50, 100, 200, 400}.
- the number of categories p is chosen among {2, 3}

For each value of the triplet of parameters, we sample 100 MR-Sort models $\mathcal{M}^0$, and record the computation time (t) needed to provide a model $\mathcal{M}^{\mathrm{NCS}}$

Fig. 3 displays the time needed by Algorithm 1 to compute $\mathcal{M}^{NCS}$, versus the number of reference alternatives $|\mathbb{X}^\star|$, both represented in logarithmic scale, in various configurations of the number of criteria. The fact that each configuration is seemingly represented by a straight line hints at a linear dependency between $\log t^{SAT}$ and $\log |\mathbb{X}^\star|$. The fact that the various straight lines, corresponding to various number of criteria, seem parallel, with a slope close to 1, is compatible to a law where t_{SAT} is proportional to $|X^\star|$. The same observations in the plane (number of criteria × computation time) (not represented) leads to infer a law

$$t^{SAT} \;\; \propto \;\; |\mathbb{X}^\star| \times 2^{|\mathcal{N}|},$$

where the computing time is proportional to the number of reference alternatives and to the number of coalitions (corresponding to the number of $|\mathcal{N}|$-ary clauses of the SAT formulation). Finally, as a rule of thumb: *the average computation time is about 10 s for 11 criteria, 3 categories and 100 reference alternatives; it doubles for each additional criterion, or when the number of reference alternatives doubles.*

5.3. Comparison between the formulations

In order to compare between models, we focus on a situation with three categories, nine criteria, and 64 reference alternatives, serving as a baseline. We then consider situations deviating from the baseline on a single parameter – either the number of categories p, from 2 to 5, or the number of criteria, among {5, 7, 9, 11, 13}, or the number of reference alternative among {16, 32, 64, 128, 256}. For each considered value of the triple of parameters, we sample 50 MR-Sort models representing the ground truth $\mathcal{M}^0$, and we record the computation time t needed to provide each of the three models $\mathcal{M}^{\mathrm{NCS}}$, $\mathcal{M}^{\mathrm{MIP-D}}$ and $\mathcal{M}^{\mathrm{MIP-O}}$, as well as the generalization indexes for the three models. The MIP are solved with a timeout of one hour.

5.3.1. Results on the computation time.

For the three formulations under scrutiny and the set of considered parameters governing the input, the computation time ranges from below the tenth of a second to an hour (when the timeout is reached), thus covering about five orders of magnitude. The left side of Fig. 4 depicts the distribution of the computation time for the baseline situation (9 criteria, 3 categories, 64 reference assignments). While the computing time for the SAT and the MIP-D formulations seem to be centered around similar values (with
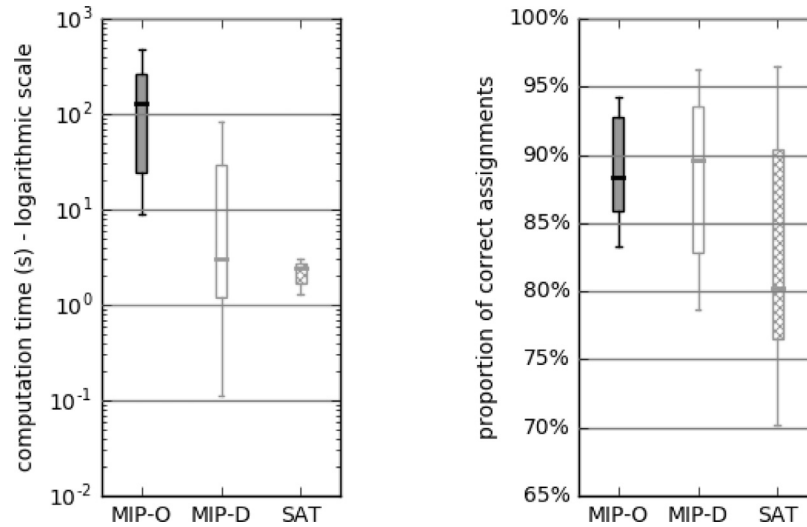
Fig. 4. Distribution of the computation time and the proportion of assignment similar to the ground truth for the three models in the baseline configuration: 9 criteria, 3 categories, 64 reference alternatives. Represented: median; box: $25 - 75\%$; whiskers: $10 - 90\%$.
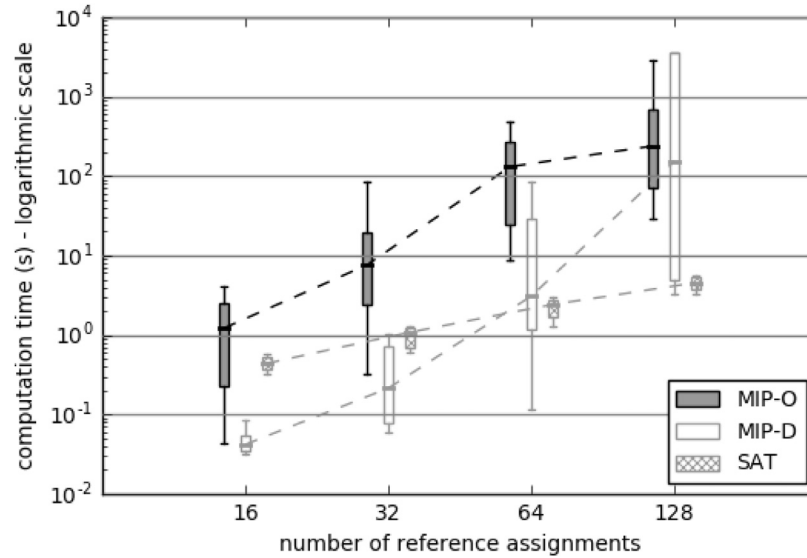
Fig. 5. Distribution of the computation time for the three models by number of reference assignments, with three classes and nine criteria.

$Med(t^{SAT}) \approx 2.4$ s and $Med(t^{MIP-D}) \approx 3.1$ s for the baseline), the distribution of the computing time for the SAT algorithm around this center is very tight, while the spread of this distribution for the MIP-D formulation is comparatively huge: The slowest tenth of instances run about a thousand time slower than the quickest tenth. The computation time of the MIP-O formulation appears about 50 times slower than the SAT, with a central value of $Med(t^{MIP-O}) \approx 130$ s, and covers about two orders of magnitude.

In order to assess the influence of the parameters governing the size and complexity of the input, we explore situations differing from the baseline on a single parameter.

- *The number of reference assignments $|\mathbb{X}^*|$.* Fig. 5 indicates that the distribution of the computing time for SAT-based algorithm remains tightly grouped around its central value, and that this value steadily increases with the number of reference assignments. Meanwhile, the two MIP formulations display a similar behavior, with an increase of the central tendency steeper than the one displayed by the SAT, and a spread that widens when taking into account additional reference assignments.

- *The number of criteria $|\mathcal{N}|$.* Fig. 6 indicates that the distribution of the computing time for SAT-based algorithm remains tightly grouped around its central value, and that this value steadily increases with the number of criteria. This increase is steeper in the case of the SAT and MIP-O formulations than for the MIP-D formulation.

- *The number of categories p.* Fig. 7 displays an interesting phenomenon. The number of categories seems to have a mild influence on the computation time, without any restriction for the SAT-based algorithm, and as soon as there are three categories or more for the MIP-based algorithm, with a clear exception in the case of two categories, which yields instances of the problem solved ten times faster than with three or more categories.

5.3.2. Results on the ability of the inferred model to restore the original one.

The right side of Fig. 4 depicts the distribution of the proportion of correct assignments (as compared to the ground truth) for the baseline situation (9 criteria, 3 categories, 64 reference assignments). The situation depicted is conveniently described by using the distribution of outcomes yielded by the MIP-D formulation as

Fig. 6. Distribution of the computation time for the three models by number of criteria, with three classes and 64 learning examples.
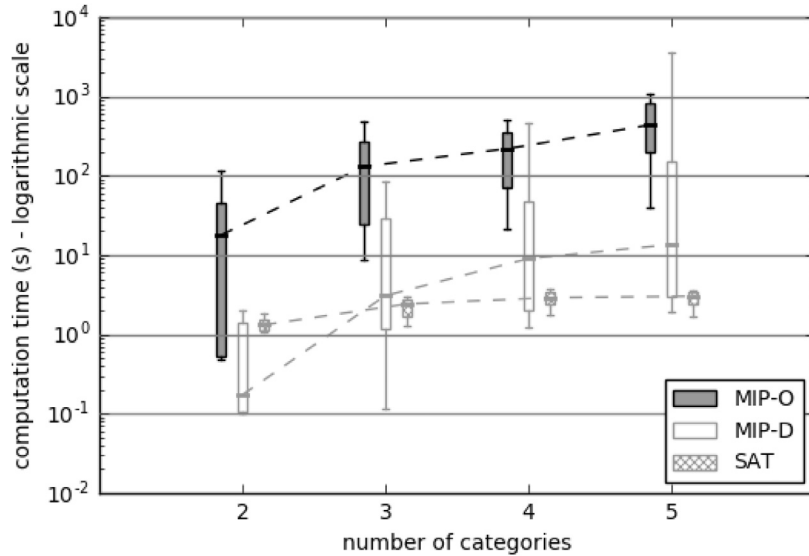
Fig. 7. Distribution of the computation time for the three models by number of categories, with nine criteria and 64 learning examples.

a pivotal point to which we compare those yielded by the SAT and MIP-O formulations: the central 80% of the distribution (between the whiskers) of outcomes for the MIP-O corresponds to the central half (the box) for the MIP-D, while the best half of the distribution of outcomes for the SAT corresponds to the central 80% for the MIP-D. In other terms, compared to the MIP-D, the MIP-O offers consistently good results, while the SAT has a 50% chance to yield a model that does not align very well with the ground truth.

Figs. 8–10 depict the variations of the alignment of the models yielded by the three algorithms with the ground truth with respect to the number of reference assignments, of criteria, or of categories, respectively. The experimental results display a tendency towards a degradation of this alignment as the number of criteria or the number of categories increase. Conversely, as expected, increasing the number of reference assignments noticeably enhances the restoration rate. The three algorithms seem to behave in a similar manner with respect to the modification of these parameters.

5.3.3. Reliability

The three formulations expressing the problem we solve – finding a non compensatory sorting model extending a given assignment of reference alternatives – into technical terms are theoretically faithful. Moreover, as we generate the input assignment with a hidden *ground truth* which itself obeys a non-compensatory sorting model, the search we set out to perform should provably succeed. Unfortunately, a computer program is but a pale reflection of an algorithm, as it is restricted in using finite resources. While we take great care in designing the experimental protocol in order to avoid memory problems, we have purposefully used off-the-shelf software with default setting to solve the formulations. While this attitude has given excellent result for the implementation of the SAT-based algorithm, which has never failed to retrieve a model that succeeds in extending the given assignment, the two MIP-based implementations have suffered from a variety of failures, either not terminating before the timeout set at one hour or wrongly concluding on the infeasibility of the MIP. We report these abnormal behaviors in Table 4.

Fig. 8. Distribution of the generalization index for the three models by size of the learning set, with three classes and nine criteria.
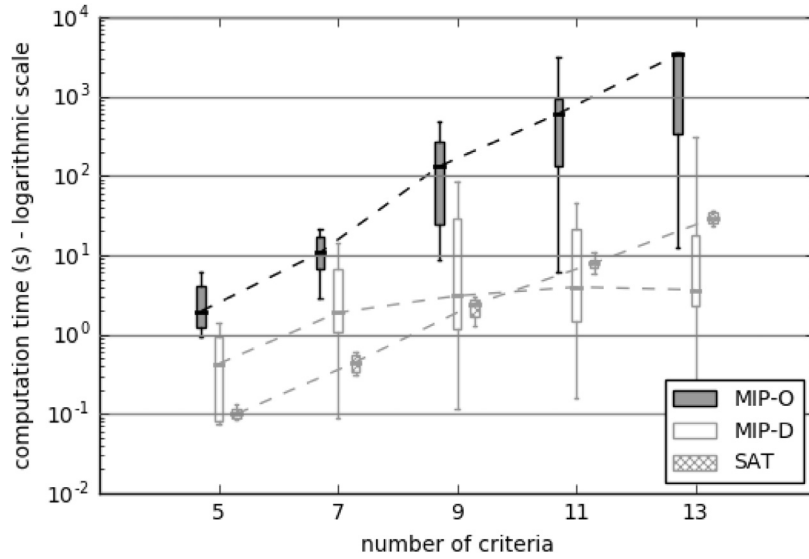

Fig. 9. Distribution of the generalization index for the three models by number of criteria, with three classes and 64 learning examples.
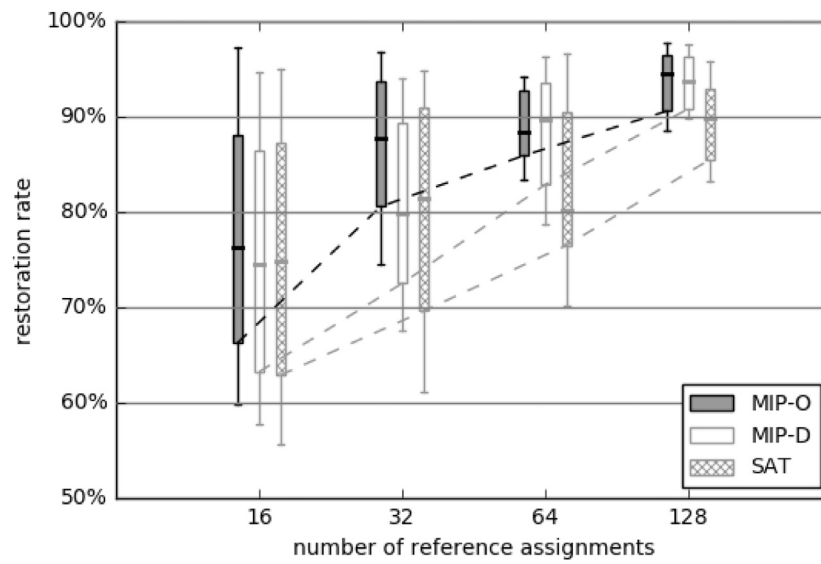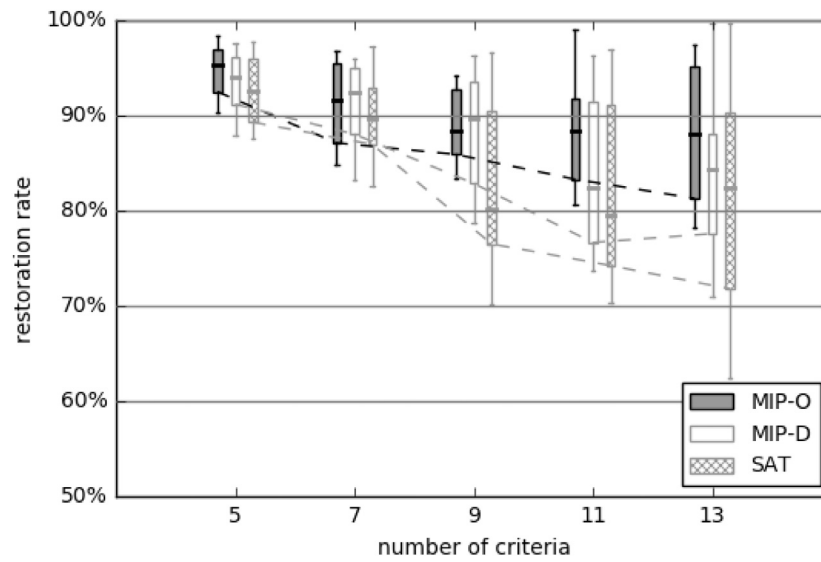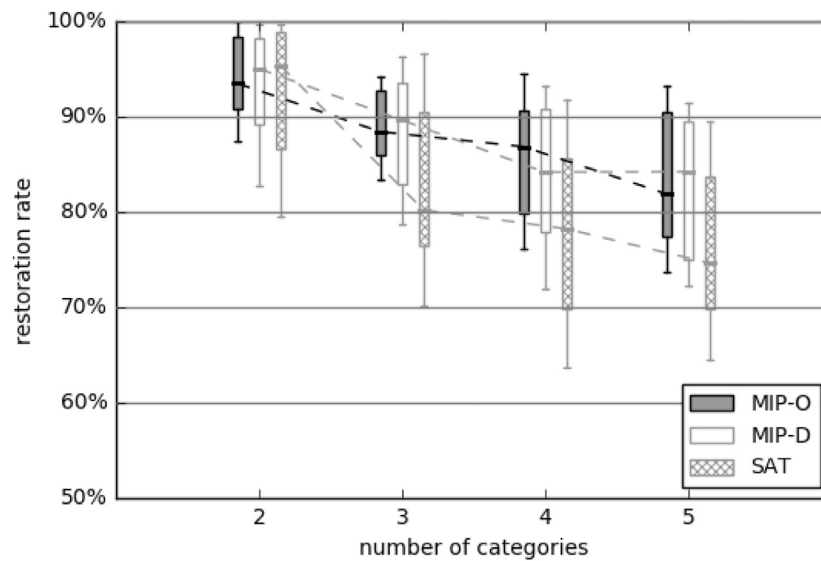

Fig. 10. Distribution of the generalization index for the three models by number of categories, with nine criteria and 64 learning examples.

Table 4
Proportion of instances failing to retrieve a model. The default case is due to reaching the time limit, except for configurations marked with a dagger where the failure is due to an alleged infeasibility of the formulation.

Number of criteria	5	7	9	11	13	9	9	9
Number of categories	3	3	3	3	3	3	5	7
Number of reference assignments	64	64	64	64	64	128	64	64
MIP-D	4%[†]	8%[†]	4%	0	0	42%	10%	12%
MIP-O	0	0	0	10%	48%	4%	0	0
SAT	0	0	0	0	0	0	0	0

6. Discussion and perspectives

In this section, we strive at interpreting the results presented in Section 5. In Section 6.1, we address the influence of the parameters governing the size and structure of the input - the reference assignment we set out to extend with a non-compensatory sorting model - on the computing time of the programs implementing the three formulations modeling the problem. In Section 6.2, we discuss the differing approaches to knowledge representation underlying these different formulations, and their practical consequences.

6.1. Influence of the parameters

The influence of the various parameters ($|\mathbb{X}^\star|$, the number of reference assignments; $|\mathcal{N}|$, the number of criteria; p, the number of categories) governing the input on the ability of the output model to predict the ground truth seeding the input is best understood from a machine learning perspective. The input assignments form the learning set of the algorithm, while the number of criteria and the number of categories govern the number of parameters describing the non-compensatory sorting model. Hence, an increase in $|\mathbb{X}^\star|$ adds constraints upon the system, while increases in $|\mathcal{N}|$ or p relieve some constraints, but demand more resources for their management.

- The comparison between MIP-O and MIP-D informs the influence of the loss function. This influence is threefold: optimizing this function demands a lot more time than simply returning the first admissible solution found; formalizing the problem of extending the input assignment with a model as an optimization problem incorporates a kind of robustness into the algorithm, which translates to a decrease in the number of failures; paradoxically, the strategy consisting in finding the most representative model (in the sense of the chosen loss function) does not yield models with a better alignment to the ground truth than the one consisting to return a random suitable model.

- The MIP-D and SAT formulations implement the same binary attitude concerning the suitability of a non-compensatory model to extend a given assignment, and both arbitrarily yield the first-encountered suitable model. Nevertheless, algorithms based on these formulations display marked differences in behavior: while the running time of the SAT-based algorithm is very homogeneous between instances and follows very regular patterns when the input parameters change, the MIP-D algorithm behaves a lot more erratically, with some failures (displayed in Table 4) and a tremendous spread. We credit this difference in behavior to a difference of approach to knowledge representation, as discussed in Section 6.2. Also, with the same input parameters, the model returned by the MIP-D algorithm seems on average to be more faithful to the ground truth than the model returned by the SAT algorithm. As both models return random suitable models in different categories (MR-Sort for the MIP, and the superset NCS for the SAT, while the ground truth is chosen in the MR-Sort category), we interpret the dif-

ference in the proportion of correct assignment to the respective volumes of the two categories of model, and discuss the pros and cons of assuming one or the other in Section 6.2.

- Reference assignments are a necessary evil. On the one hand, they provide the information needed to entrench the model, and refine the precision up to which its parameters can be known. On the other hand, they erect a computational barrier which adds up more quickly for the MIP formulations we are considering than for the SAT one, as shown in Fig. 5. Overcoming this barrier demands time and threatens the integrity of the somewhat brittle numerical representation underlying the MIP-D formulation.

- From the perspective of the model-fitting algorithm, the number of criteria and the number of categories are usually exogenous parameters, fixed according to the needs of the decision situation. The specific numbers of criteria we considered during the experiment, from 5 to 13, cover most of the typical decision situations considered in MCDA. Introducing more criteria demands to assess more parameters, which has a compound effect on complexity, as it requires at the same time to build a higher dimension representation of the models, and to provide more reference examples in order to be determined with a precision suitable to decision making. Apart from a noticeable exception (see below), the number of categories does not seem to have much influence (as shown on Figs. 7 and 10).

- Underconstrained models are not very good at providing recommendations. When fed with scarce information, the task of finding a suitable extension is easy, but there are very little guarantees this extension matches the unexpressed knowledge and preferences of the decision maker concerning alternatives outside the learning set. We interpret the decrease in the ability to align with the ground truth as the number of criteria increases displayed on Fig. 9 as an expression of an *overfitting* phenomenon, where too many parameters are chosen to faithfully represent a too little slice of the set of alternatives, but poorly represent cases never seen before.

- Mixed integer programs can represent decision problems, in theory. Practically though, some complex inputs have proven overwhelming for the MIP-D formulation, whereas the MIP-O has shown more robustness, as evidenced by Table 4. It seems fair to assume this lack of stability is related to the absence of a normalization constraint such as (3d) in the MIP-D formulation . Determining a good lower bound on the difference of normalized voting power between sufficient and insufficient coalitions would therefore likely help alleviating this issue.

- MR-Sort with two categories is structurally different than models with more than two categories. While we have defined it as a procedure where alternatives are compared holistically to a profile, it can also be described as an additive value sorting model with stepwise, non-decreasing, 2-valued marginals. The experimental results, both for the computing time and the alignment with the ground truth (see Figs. 7 and 10, where the points corresponding to two categories are outliers with respect to the rest of the series) highlight this peculiarity, and tend to

show that the value-based representation of the MR-Sort model with two categories is computationally efficient.

6.2. Numeric or symbolic representation of coalitions

Our proposal to infer non-compensatory sorting models from assignment examples using a SAT formulation relies on a symbolic representation of sufficient coalitions of criteria. It departs fundamentally from the state-of-the-art approach of representing the upset of sufficient coalitions with a numeric majority rule (MR) .

Obviously, U-NCS is more general than MR-Sort as additive weights/majority level induce a set of minimal criteria coalitions, while a set of minimal coalitions might not be additive. Uyanik et al. (2017) studies the proportion of additive representations: all NCS models are additive up to 3 criteria, but the proportion of additive NCS models tends to be quickly marginal when the number of criteria increases. It is also possible to extend the MR-Sort model up to U-NCS by considering a capacity instead of a weight vector (see e.g. Sobrie et al., 2015). This leads to MIP formulations of increasing computational difficulty as the arity of the capacity increases (increase of the number of decision variables and the computation time). Also, (Sobrie et al., 2015) shows that a MR-Sort model learned from NCS generated examples provides a good approximation of this NCS model.

A distinctive feature of MR-Sort is its parsimony with respect to interaction between criteria, a notion that the SAT formulation of U-NCS fails to capture. However, there are many ways to additively represent a set of minimal coalitions, and the intuitive interpretation of the weights can therefore be misleading: there is no one to one correspondence between the tuples of voting powers and majority level, and the sets of additive coalitions of criteria. For instance, consider a three criteria problem in which coalitions of criteria are sufficient if and only if their cardinality is at least two. This set of minimal coalitions can be represented by $w = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\lambda = \frac{1}{2}$, or $w = (0.49, 0.49, 0.02)$ and $\lambda = \frac{1}{2}$. It is obvious that these two numerical representations yield erroneously to two very distinct interpretations about the relative importance of criteria. In this sense, the symbolic representation avoiding weights used in the SAT formulation is more faithful than a numerical representation. As a consequence, this non-uniqueness of the additive representation penalizes the effectiveness of loss functions involved in MIPs.

Also, as mentioned in Section 5.3, the feasibility version of the MIP suffers from numerical instability, perhaps because of the lack of a normalization constraint. The symbolic representation of sufficient coalitions circumvents the difficult mathematical question of providing a good lower bound on the worst case difference of normalized voting power between sufficient and insufficient coalitions.

7. Conclusion

In this paper, we consider the multiple criteria non-compensatory sorting model (Bouyssou and Marchant, 2007a; 2007b) and propose a new SAT formulation for inferring this sorting model from a learning set provided by a DM. Learning this model has already been addressed by the literature, and solved by the resolution of a MIP (Leroy et al., 2011) or via a specific heuristic (Sobrie et al., 2013; 2015). Due to a high computation time, the MIP formulation can only apply to learning sets of limited size. Heuristic methods can handle large datasets, but can not ensure to find a compatible model with the learning set whenever it exists. Our new algorithm provides such guaranty. We implemented and tested our SAT formulation, and it outperforms MIP approaches in terms of computation time (reduction of the computation time by a factor of about 50).

Moreover, it could have been the case that this good performance in terms of computing time would be counterbalanced by a limited ability of the inferred model in terms of generalization. Indeed, a MIP approach focuses the effort in finding a relevant representative model among the compatible models (through the use of an objective function), while our SAT approach does return the first compatible model found.

Our experiments show that MIP and SAT approaches have similar performances in terms of generalization. Therefore, we believe this algorithm to be a strong advance in terms of learning NCS models based on learning sets, in particular when learning sets become relatively large.

Thanks to its efficiency – finding a model compatible to some preference information takes seconds instead of minutes – this algorithm is well suited to be embedded in an interactive process, where the decision maker is invited to interactively elicit a non-compensatory sorting model by incrementally building a learning set (and possibly additional preference information). Another line of research lies in the idea of using the feasible SAT solution as a "warm start" to improve the resolution of the MIP-O formulation.

In order to address real-world decision aiding situations, the algorithm we propose needs to be equipped with techniques permitting to account for noisy or inconsistent data. While the numeric formulations may rely on Lagrangian techniques to handle the requirement of correctly representing the data as a set of soft constraints rather than hard ones, the logic formulation we propose could usefully investigate the notions of maximally consistent or minimally inconsistent set of clauses (see e.g. Besnard et al., 2015 for solving techniques, or e.g. (Mousseau et al., 2003) for an application in a MCDA context). The increased speed, as compared to the previous MIP-based approach, opens the door to the exploration of the set of all U-NCS models extending a given assignment, in the vein of the version space theory (Mitchell, 1982) and robust decision aiding (Greco et al., 2008). The knowledge representation underlying our approach may also permit to support a recommendation with an *explanation* (Amgoud and Serrurier, 2008; Belahcène et al., 2017a; Labreuche, 2011).

References

Amgoud, L., Serrurier, M., 2008. Agents That Argue and Explain Classifications. In Autonomous Agents and Multi-Agent Systems. AAMAS.

Belahcène, K., Labreuche, C., Maudet, N., Mousseau, V., Ouerdane, W., 2017. A model for accountable ordinal sorting. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. IJCAI, pp. 814–820.

Belahcène, K., Mousseau, V., Pirlot, M., Sobrie, O., 2017. Preference Elicitation and Learning in a Multiple Criteria Decision Aid Perspective. LGI Research report 2017-02, CentraleSupélec. http://www.lgi.ecp.fr/Biblio/PDF/CR-LGI-2017-02.pdf

Besnard, P., Grégoire, E., Lagniez, J.M., 2015. On computing maximal subsets of clauses that must be satisfiable with possibly mutually-contradictory assumptive contexts. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence January 25–30. Austin, Texas, USA.

Biere, A., Heule, M., van Maaren, H., Walsh, T., 2009. Handbook of Satisfiability. Frontiers in Artificial Intelligence and Applications 185. IOS Press.

Bouyssou, D., Marchant, T., 2007. An axiomatic approach to non compensatory sorting methods in MCDM, i: the case of two categories. Eur. J. Oper. Res. 178 (1), 217–245.

Bouyssou, D., Marchant, T., 2007. An axiomatic approach to noncompensatory sorting methods in MCDM, II: more than two categories. Eur. J. Oper. Res. 178 (1), 246–276.

Bouyssou, D., Marchant, T., Pirlot, M., Tsoukiàs, A., Vincke, P., 2006. Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst. International Series in Operations Research and Management Science, Volume 86, First edition Springer, Boston.

Cook, S., 1971. The complexity of theorem proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing, pp. 151–158.

Dickerson, J., Goldman, J., Karp, J., Procaccia, A., Sandholm, T., 2014. The computational rise and fall of fairness. AAAI 1405–1411.

Fürnkranz, J., Hüllermeier, E., 2011. Preference Learning. Springer-Verlag New York, Inc.

Greco, S., Kadzinski, M., Słowinski, R., 2011. Selection of a representative value function in robust multiple criteria sorting. Comput. Oper. Res. 38 (11), 1620–1637.

Greco, S., Matarazzo, B., Słwinski, R., 2002. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. Eur. J. Oper. Res. 138 (2), 247–259.

Greco, S., Matarazzo, B., Słowinski, R., 2016. Decision Rule Approach. In: Greco, S., Ehrgott, M., Figueira, J. (Eds.), in multiple criteria decision analysis - State of the art surveys, pp. 497–552.

Greco, S., Mousseau, V., Słowinski, R., 2008. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. Eur. J. Oper. Res. 191 (2), 416–436.

Greco, S., Mousseau, V., Słowinski, R., 2010. Multiple criteria sorting with a set of additive value functions. Eur. J. Oper. Res. 207 (3), 1455–1470.

Jacquet-Lagrèze, E., Siskos, Y., 1982. Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. Eur. J. Oper. Res. 10 (2), 151–164.

Labreuche, C., 2011. A general framework for explaining the results of a multi-attribute preference model. Artif. Intell. 175 (7–8), 1410–1448.

Leroy, A., Mousseau, V., Pirlot, M., 2011. Learning the Parameters of a Multiple Criteria Sorting Method. In: Brafman, R., Roberts, F., Tsoukiás, A. (Eds.), Algorithmic Decision Theory, volume 6992 of Lecture Notes in Artificial Intelligence. Springer, pp. 219–233.

Marichal, J.L., Meyer, P., Roubens, M., 2005. Sorting multi-attribute alternatives: the TOMASO method. Comput. Oper. Res. 32 (4), 861–877.

Meyer, P., Olteanu, A.L., 2017. Integrating large positive and negative performance differences into multicriteria majority-rule sorting models. Comput. Oper. Res. 81, 216–230.

Mitchell, T., 1982. Generalization as search. Artif. Intell. 18 (2), 203–226.

Moskewicz, M., Madigan, C., Zhao, Y., Zhang, L., Malik, S., 2001. Chaff: Engineering an Efficient SAT Solver. In: In Proceedings of the 38th annual Design Automation Conference (DAC '01). ACM, New York, NY, USA, pp. 530–535.

Mousseau, V., Figueira, J., Dias, L.C., Silva, C.G.d., Clímaco, J.C.N., 2003. Resolving inconsistencies among constraints on the parameters of an MCDA model. Eur. J. Oper. Res. 147 (1), 72–93.

Mousseau, V., Słowinski, R., 1998. Inferring an ELECTRE TRI model from assignment examples. J. Global Optim. 12 (2), 157–174.

Roy, B., Bouyssou, D., 1993. Aide multicrità la décision: Méthodes et cas. Economica, Paris.

Sobrie, O., Mousseau, V., Pirlot, M., 2013. Learning a Majority Rule Model from Large Sets of Assignment Examples. In: Perny, P., Pirlot, M., Tsoukiàs, A. (Eds.), Algorithmic Decision Theory, pages 336–350. Springer, Brussels, Belgium.

Sobrie, O., Mousseau, V., Pirlot, M., 2015. Learning the Parameters of a Non Compensatory Sorting Model. In: Algorithmic Decision Theory, ADT 2015, pp. 153–170.

Soos, M., 2016. The cryptominisat 5 set of solvers at SAT competition 2016. In: In Proceedings of SAT Competition 2016: Solver and Benchmark Descriptions. University of Helsinki. Volume B-2016-1 of Department of Computer Science Series of Publications B

Soylu, B., 2011. A multi-criteria sorting procedure with Tchebycheff utility function. Comput. Oper. Res. 38 (8), 1091–1102.

Uyanik, E.E., Sobrie, O., Mousseau, V., Pirlot, M., 2017. Enumerating and categorizing positive boolean functions separable by a k-additive capacity. Discrete Appl. Math. 229, 17–30.

Zheng, J., Metchebon Takougang, S.A., Mousseau, V., Pirlot, M., 2014. Learning criteria weights of an optimistic electre tri sorting rule. Comput. Oper. Res. 49, 28–40.

Zopounidis, C., Doumpos, M., 2002. Multicriteria classification and sorting methods: a literature review. Eur. J. Oper. Res. 138 (2), 229–246.

CrossMark

Explaining robust additive utility models by sequences of preference swaps

K. Belahcene[1] · C. Labreuche[2] · N. Maudet[3] ·
V. Mousseau[1] · W. Ouerdane[1]

Abstract As decision-aiding tools become more popular everyday—but at the same time more sophisticated—it is of utmost importance to develop their explanatory capabilities. Some decisions require careful explanations, which can be challenging to provide when the underlying mathematical model is complex. This is the case when recommendations are based on incomplete expression of preferences, as the decision-aiding tool has to infer despite this scarcity of information. This step is key in the process but hardly intelligible for the user. The robust additive utility model is a necessary preference relation which makes minimal assumptions, at the price of handling a collection of compatible utility functions, virtually impossible to exhibit to the user. This strength for the model is a challenge for the explanation. In this paper, we come up with an explanation engine based on sequences of preference swaps, that is, pairwise comparison of alternatives. The intuition is to confront the decision maker with "elementary" comparisons, thus building incremental explanations. Elementary

✉ N. Maudet
nicolas.maudet@lip6.fr

K. Belahcene
khaled.belahcene@centralesupelec.fr

C. Labreuche
christophe.labreuche@thalesgroup.com

V. Mousseau
vincent.mousseau@centralesupelec.fr

W. Ouerdane
wassila.ouerdane@centralesupelec.fr

[1] LGI, CentraleSupélec, Université Paris-Saclay, Chatenay Malabry, France

[2] Thales Research & Technology, 91767 Palaiseau Cedex, France

[3] Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 75005 Paris, France

here means that alternatives compared may only differ on two criteria. Technically, our explanation engine exploits some properties of the necessary preference relation that we unveil in the paper. Equipped with this, we explore the issues of the existence and length of the resulting sequences. We show in particular that in the general case, no bound can be given on the length of explanations, but that in binary domains, the sequences remain short.

1 Introduction

A decision-aiding problem consists in formalizing the problem and eliciting the preferences of the decision maker (DM) to make recommendations. In many decision contexts, only providing recommendations based on the elicited preference model is insufficient. In fact, decision makers may want explanations which justify in a convincing way such recommendations. Indeed, justifying and explaining a rationale for a decision is almost as important as the recommendation itself. Building a convincing explanation is often required when the DM cannot be assumed to have any mathematical background, as in the case of online recommender systems, where it has been shown that explanations improve the acceptability of the recommended choice (Pu and Chen 2007; Symeonidis et al. 2009; O'Sullivan et al. 2007). But even experts of a domain can have huge difficulty to grasp with the mathematical models underlying some decision-aiding tools. In this case, it is not satisfactory to just put forward the preference model and the resulting recommendation. Although technically, of course, this model does contain all the information on which the recommendation is based, the format is unlikely to be suitable for presentation. Hence, the need for a synthetic, short and easy to understand explanation.

Depending on the setting considered, the nature of an explanation may greatly vary. Sometimes, even vague statements can prove effective to persuade a specific decision maker. But when the decision is important, or when the decision maker is accountable for the decision chosen (a situation where the decision needs to be justified to some other stakeholders who did not participate to the decision process), the explanation should be viable even under close scrutiny. Complete explanations provide some guarantees in that respect since they bring all the information required to reconstruct the rationale of the recommendation—in a sense they formally "prove" it.

In this paper, we shall thus concentrate on complete explanations in the context of decisions involving multiple criteria. More precisely, we propose to construct pieces evidence that support unambiguously a binary preference relation between two alternatives described along multiple attributes. Such a relation is very often not explicit but elicited by some algorithmic process from preference information stated by the decision maker. In our case, this initial information takes the form of pairwise comparisons of alternatives. This initial input may be scarce, in any case not sufficient to fully specify the preference relation of the DM. To deal with the incompleteness of the

expression of preferences, the decision-aiding method will make use of an inference step. It is usually an involved process, challenging for explanation.

Our explanation engine takes inspiration from the even-swaps method (Hammond et al. 1998), an elicitation procedure assuming an additive value model of preferences and based on trade-offs between pairs of attributes (hence the name even swaps). Broadly speaking, in each swap, the DM changes the score of an alternative on one attribute, and compensates this change with one another attribute, so that the new alternative is equally preferred. The process is repeated until dominance can be shown to hold, allowing to progressively eliminate attributes. The idea is to use similar sequences as explanations of a recommendation. The problem with such a process is that it requires each new generated option to be equally preferred to the initial one, which is poorly adapted to the context of incomplete preferences (as such an equivalence virtually never holds). To circumvent this issue, we propose a generalization of even swaps to preference swaps, and simply exhibit a comparison between alternatives. To keep the sequence as simple as possible, we aim at constructing a sequence of low-order preference swaps between two alternatives, in the sense that two successive alternatives in the sequence only differ on a few criteria. In the end, the resulting explanations can be appreciated through the number of swaps (length of an explanation) and the order of the most complex swap involved in the explanation (the number of differing attributes between the two alternatives). An interesting feature of this explanation engine is that it can be shown to operate on any value-based decision models satisfying some basic axiomatic properties.

We propose thereafter to instantiate the engine by relying on a robust additive utility model (Greco et al. 2008, 2010). The robust (necessary) relation is constructed according to preference information provided by the decision maker. However, contrary to the classical additive models, in the robust approach the relation holds if any possible completion of the available preferential information yields the preferential statement. In fact, in additive models, such as UTA (Additive UTility) (Jacquet-Lagrèze and Siskos 1982), the preferential information brought by the DM is not sufficient to uniquely specify the utility functions (utility functions are only partially known), but the multiplicity of the compatible utility is not taken into account. To provide a solid mechanism to construct explanations for necessary preference relations, we come up with a new characterization of the necessary preference relation, based on the notion of covectors, that facilitates its implementation in the explanation engine.

In a nutshell, our proposal is thus to decompose a robust preference into several simpler recommendations. This paper investigates this idea and tackles the following questions: are such explanations guaranteed to exist, in particular if we restrict the order of swaps? And if they do exist, can we exhibit upper bounds on their length? As we shall see, the answer to this question crucially depends on the number of distinct values referenced by the preference information. In binary domains, we provide an efficient algorithm which constructs such explanations.

The remainder of the paper is as follows. Section 2 presents the explanation engine which relies on the construction of sequence of preference swaps between two alternatives. In Sect. 3, we define and analyze the value-based robust preference relation. Section 4 proposes results concerning the construction of explanations when preference information is expressed using two levels on each criterion. Finally, Sect. 5

studies how our contributions relate to previous work and proposes extensions and further work.

2 The explanation engine

2.1 Presentation of the decision context

This article is set in the context of Multicriteria Decision Making, where a decision maker has to decide between several alternatives explicitly measured on several criteria. We call N the set of criteria, so alternatives are represented by elements of a set $\mathbb{X} = \prod_{i \in N} \mathbb{X}_i$, where the attribute set $\mathbb{X}_i$ for criterion $i \in N$ is totally ordered by the relation $\succsim_i$ denoting preference.

Example 1 You need to chose a hotel for a business trip, and you are undecided between four options described by the performance table (see below). Such options are evaluated according to four criteria.

- The room comfort, ranging from $*$ (low) to $*****$ (high).
- The presence of a restaurant on the premise, with yes preferred to no.
- The commute time to the convention center, the lower the better.
- The cost, the lower the better.

Hotel	Comfort	Restaurant	Commute time (min)	Cost
h_1	$5*$	Yes	10	160 \$
h_2	$4*$	Yes	45	180 \$
h_3	$3*$	No	15	60 \$
h_4	$2*$	No	60	50 \$

Definition 1 (*ceteris paribus sets of pairs of alternatives*) for any partition of criteria $N = A \cup (N \backslash A)$ and corresponding partition of attributes $x_A \in \prod_{i \in A} \mathbb{X}_i$ and $x_{-A} \in \prod_{i \notin A} \mathbb{X}_i$, (x_A, x_{-A}) is an alternative belonging to $\mathbb{X}$. For $x_A, y_A \in \prod_{i \in A} \mathbb{X}_i$, we define the ceteris paribus set $(x_A, y_A)_{cp}$ as the set of every possible completions of the pair:

$$(x_A, y_A)_{cp} := \left\{ ((x_A, c_{-A}), (y_A, c_{-A})), c_{-A} \in \prod_{i \notin A} \mathbb{X}_i \right\}$$

When comparing two alternatives, the criteria may unanimously rank one alternative above the other.

Definition 2 (*weak Pareto dominance*)

$$\forall (x, y) \in \mathbb{X} \times \mathbb{X}, \quad (x, y) \in \mathcal{D} \iff \forall i \in N, \quad x_i \succsim_i y_i$$

Definition 3 (*sets of shared and differing attributes*)

$$\forall x, y \in \mathbb{X}, N^=_{(x,y)} := \{i \in N : x_i = y_i\} \quad \text{and} \quad N^{\neq}_{(x,y)} := \{i \in N : x_i \neq y_i\}$$

Preferences of the decision maker make up a binary relation between alternatives $\mathcal{R} \subset \mathbb{X}^2$, so that $(x, y) \in \mathcal{R}$ denotes the (weak) preference of alternative x over alternative y. More often than not, this relation is not explicit over $\mathbb{X}^2$, but elicited, extrapolated by some algorithmic process from preference information stated by the decision maker. In this context, an explanation of a statement $(x, y) \in \mathcal{R}$ is a piece of supportive evidence, enabling the decision maker to assert this preference. The explanation engine we develop in Sect. 4 assumes the relation $\mathcal{R}$ satisfies three core axioms:

Axiom 1 (compatibility to dominance) $\mathcal{D} \subset \mathcal{R}$
Axiom 2 (transitivity) $\forall x, y, z \in \mathbb{X} : (x, y) \in \mathcal{R} \wedge (y, z) \in \mathcal{R} \Rightarrow (x, z) \in \mathcal{R}$
Axiom 3 (cancelation) *For any ceteris paribus set of pairs s, if a pair of alternatives in s is in relation $\mathcal{R}$, then every pair of alternatives in s is in relation $\mathcal{R}$.*

Example 2 (Ex. 1 cont.) Hotel h_1 dominates hotel h_2, as it is at the same time more comfortable, closer to the convention center, and cheaper, while being as good on the criterion presence of a restaurant. Thus, $(h_1, h_4) \in \mathcal{D}$, and $(h_1, h_4) \in \mathcal{R}$.

Hotels h_3 and h_4 share their absence of a restaurant on the premise. Thus, preference of one over the other ignores the criterion restaurant and is represented by the ceteris paribus set $\left((3*, \_\_r, 15\text{min}, 60\$), (2*, \_\_r, 60\text{min}, 50\$)\right)_{cp}$, where $\_\_r$ stands for any value in $\mathbb{X}_r$. As $\mathbb{X}_r$ contains two distinct values, there are two pairs in this set, and $(h_3, h_4) \in \mathcal{R} \iff \left((3*, \text{yes}, 15\text{min}, 60\$), (2*, \text{yes}, 60\text{min}, 50\$)\right) \in \mathcal{R}$.

Compatibility to dominance is a fundamental requirement to correctly model preference. Transitivity asks for the model to eschew Condorcet's paradox and to behave like a preorder relation. Cancelation implies the preferential independence of criteria, so that only differing attributes have a say in determining preference.

Many popular, value-based decision models fulfill these requirements, measuring the fitness of an alternative by combining its attributes in a single index, using the average, or weighted average of the attributes, or some carefully chosen separable, parametric value function of the attributes. So does the robust additive model, described in Sect. 3.

2.2 Sequences of low-order preference swaps

The explanation engine detailed in what follows is reminiscent of the even-swaps method (Hammond et al. 1998), an interactive and constructive elicitation procedure assuming an additive value model of preferences. This method aims at identifying, between two options x and y, which one is preferred to the other, without explicitly constructing the utility functions. This is basically an elimination process based on trade-offs between pairs of attributes ("swaps"), that can be seen as a scattered exploration of the iso-preference curve of the decision maker (the curve where lies, even

virtually, the alternatives equally preferred).[1] Broadly speaking, in such a swap, the decision maker changes the consequence (or score) of an alternative on one attribute, and is asked to compensate for this change by acting on another attribute, so that the new alternative is equally preferred in the end ("even"). This creates a new fictitious alternative, that is indifferent to the previous one, with revised consequences. By replacing one option (say x) with a different but equally preferred one, the hope is that dominance will occur over y. The process is thus repeated allowing to progressively cancel irrelevant attributes, until dominance can be shown to hold, and building a sequence $x \sim e_1 \sim e_2 \ldots \sim e_{n-1}$, so that either $(e_{n-1}, y) \in \mathcal{D}$ or $(y, e_{n-1}) \in \mathcal{D}$.

Considered through the prism of explanation, even swaps have several very attractive features.

- Each swap involves only attributes on two criteria.
- The method entirely references alternatives inside the decision space $\mathbb{X}$, but not artifacts of the underlying decision model (such as utility functions), or relations between criteria.

However, the even-swaps approach suffers from a severe limitation, as it requires each new generated option to be equally preferred to the initial one. This is a steep requirement, for several reasons.

- Indifference requires compensation between criteria (Krantz et al. 1971), barring the possibility that some difference in attributes on one criterion could be impossible to compensate for.
- Indifference requires solvability of the attribute scales (Krantz et al. 1971), which naturally occurs on continuous scales but rarely between discrete ones.
- Indifference imposes a high cognitive workload on the decision maker, as it repeatedly asks for cardinal information.
- Indifference is hardly a robust notion, especially in the context of incomplete preferences.[2]

Consequently, we propose a generalization of even swaps that avoids these issues, while retaining their simplicity and being well suited to the context of incomplete preference. In preference swaps, the assumption of indifference between consecutive alternatives in the sequence $e_0 := x, e_1, \ldots, e_n := y$ is relaxed and replaced by an assumption of (weak) preference: $(e_{j-1}, e_j) \in \mathcal{R}$. The following definitions extend the notion of swaps to pairs of alternatives differing on more than two criteria.

---

[1] Equally preferred, or indifferent, alternatives are pairs in the symmetric part of the relation $\mathcal{R} : \forall x, y \in \mathbb{X}, x \sim y \iff \{(x, y), (y, x)\} \subset \mathcal{R}$.

[2] We note that [MH07,MH05] also propose to enrich the original even swaps method in a way that accounts for incomplete knowledge about the value function. They consider a "practical dominance" notion when the value of an alternative is at least as high as the value of another one with every feasible combination of parameters, this perspective being very close to the one developed in [GMS08] (see next section). However, this notion is only used for pre-processing dominated alternatives, and not integrated in the swap process, let alone used for explanatory purposes.

Definition 4 (*preference swaps of order k*)

$$\forall k \in \mathbb{N}^\star, \Delta_k = \begin{cases} \mathcal{D}, & if\,k = 1 \\ \{(x, y) \in \mathcal{R}\backslash\mathcal{D}, |N^{\neq}_{(x,y)}| = k\}, & if\,k > 1 \end{cases}$$

This definition leverages two properties assumed for the relation $\mathcal{R}$. As $\mathcal{D} \subset \mathcal{R}$ (Axiom 1), $\mathcal{R} = \bigcup_{k \leq |N|} \Delta_k$: any pair in $\mathcal{R}$ is a swap, and we try to reflect its cognitive difficulty, in the context of explanation, by its order, the lower, the simpler. Dominance relations are deemed to be simple, and are given the lowest order. For relations requiring trade-offs between criteria, we define the order of a swap as the number of differing attributes between the two alternatives.

We can now define the notion of explanation by a sequence of preference swaps. This type of explanation transforms one single preference statement $(x, y) \in \mathcal{R}$ that the decision maker needs to understand to a sequence of several preference statements $(e_{j-1}, e_j) \in \mathcal{R}$. The idea is that the initial preference (x, y) is complex to understand as the values of x and y differ on most (if not all) attributes, whereas each intermediate comparison (e_{j-1}, e_j) is much easier to understand as it involves alternatives differing only on a few attributes.

Definition 5 (*Explanation by preference swaps, order and length*) $\forall(x, y) \in \mathbb{X}^2, n \in \mathbb{N}$, an explanation of length n of the pair (x, y) for the relation $\mathcal{R}$ is a tuple $(e_0, e_1, \ldots, e_n) \in \mathbb{X}^n$ such that $e_0 = x$, $e_n = y$ and $\forall j \in \mathbb{N} : 1 \leq j \leq n$, $(e_{j-1}, e_j) \in \mathcal{R}$. The order of such explanation is the integer $k = \max\{k \in \mathbb{N} : \exists(j \in \mathbb{N} : 1 \leq j \leq n), (e_{j-1}, e_j) \in \Delta_k\}$.

As $\mathcal{R}$ is transitive (axiom 2), an explanation of a pair of alternatives is a proof that this pair belongs to $\mathcal{R}$. One can note that somehow we have two elements to appreciate the quality of the explanation. First, the number of comparisons (swaps) used to construct such an explanation. Second, its complexity which is defined by the most complex or difficult swap (with the highest order).

However, an important question regarding a pair $(x, y) \in \mathbb{X}^2$ is whether it is possible to find an explanation by preference swaps of the pair (x, y). The answer obviously depends on the bound, if any, placed upon the order of the swaps linking the explanation chain, or the length of the explanation chain. In this article, we address this issue by first putting a cap on the order (the order of an explanation being the order of its most difficult link), then looking for the possibility of finding an explanation subject to this order constraint. Then, if explanations are available, we look for short ones.

Definition 6 (*pairs explainable by low-order preference swaps*) $\forall k \in \mathbb{N}$, $\mathcal{E}_k(\mathcal{R})$ is the set of pairs $(x, y) \in \mathbb{X}^2$ for which there exists an explanation of any length and of order at most k.

There is a trade-off between the value of the cap placed upon the order of explanations and the set of pairs we are able to explain.

Theorem 1 *(hierarchy of binary relations)*

$$\mathcal{D} = \mathcal{E}_1(\mathcal{R}) \subseteq \mathcal{E}_2(\mathcal{R}) \subseteq \cdots \subseteq \mathcal{E}_k(\mathcal{R}) \subseteq \cdots \subseteq \mathcal{E}_{|N|}(\mathcal{R}) = \mathcal{R}$$

Proof – For any $(x, y) \in \mathcal{E}_1(\mathcal{R})$, there is a tuple $(e_0, e_1, \ldots, e_n) \in \mathbb{X}^n$ such that $e_0 = x, e_n = y$ and $\forall j \in \mathbb{N} : 1 \leq j \leq n$, $(e_{j-1}, e_j) \in \mathcal{D}$. As relation $\mathcal{D}$ is transitive, $(x, y) \in \mathcal{D}$, hence $\mathcal{D} \supseteq \mathcal{E}_1(\mathcal{R})$. Conversely, the sequence $e_0 := x, e_1 := y$ is an explanation of length one and of order one of any pair $(x, y) \in \mathcal{D}$, hence $\mathcal{D} \subseteq \mathcal{E}_1(\mathcal{R})$. Finally, $\mathcal{D} = \mathcal{E}_1(\mathcal{R})$.
- For $k' \geq k$, an explanation of order at most k is also an explanation of order at most k', so $\mathcal{E}_k(\mathcal{R}) \subseteq \mathcal{E}_{k'}(\mathcal{R})$.
- The sequence $e_0 := x, e_1 := y$ is an explanation of length one and of order $|N^{\neq}_{(x,y)}|$ of any pair $(x, y) \in \mathcal{R}$. As $|N^{\neq}_{(x,y)}| \leq |N|$, $\mathcal{R} \subseteq \mathcal{E}_{|N|}(\mathcal{R})$. Conversely, an explanation (of any order and any length) of a pair (x, y) is a proof by transitivity of $(x, y) \in \mathcal{R}$, thus $\mathcal{R} \supseteq \mathcal{E}_{|N|}(\mathcal{R})$. Finally, $\mathcal{R} = \mathcal{E}_{|N|}(\mathcal{R})$.   □

2.3 Some technical challenges with explanation

In this section, we highlight a number of key issues affecting the feasibility (from a theoretical, algorithmic point of view), and the satisfaction of the decision maker, recipient of the explanation (from a practical point of view): the existence, or not, of an explanation, its length and the values of the attributes referenced in the sequences. In fact, throughout this work we investigate the conditions (in terms of order of swaps) under which an explanation may exist. Moreover, we show also that the length of an explanation depends on the number of values of the attributes in the sequence (see Sect. 4 for the binary case). However, many other interesting questions related to these issues remain open and are not addressed in this paper (see Sect. 5).

- *Existence of an explanation* The first point to consider in the construction of an explanation is to make sure there is one to be found. Without any additional assumption, for a low cap k placed upon the order, it is quite possible that there are some statements that cannot be explained by preference swaps of order at most k. Technically, checking if we can explain a statement (x, y) in $\mathcal{E}_k(\mathcal{R})$, can be seen as determining if the vertices x and y are connected in the directed graph of the relation $\bigcup_{1 \leq n \leq k} \Delta_n$. Of course, we have efficient algorithms to test if a graph is connected or not (Even and Tarjan 1975). However, it may be challenging to use them with regard to the size of the graph (possibly infinite, and, when finite, exponential in the number of criteria) in our context.
- *Length of an explanation* A second point that we address here is the length n of the sequence. Indeed, keeping the explanation short has a great bearing on its ability to convince. Even if each elementary comparison $(e_{j-1}, e_j) \in \mathcal{R}$ is trivial for the decision maker, the overall sequence $(x, e_1, \ldots, e_{n-1}, y)$ cannot be seen as a convincing explanation if it is too long. One then looks for the shortest possible explanations, and hope for an upper bound on this minimal size. Finding the shortest explanation means resolving the problem of shortest path in the directed graph $\bigcup_{1 \leq n \leq k} \Delta_n$. Thus, the length of a shortest explanation is bounded by the

diameter of this graph.[3] Finding such a diameter is a classical problem in graph theory for which we have polynomial algorithm in terms of number, if finite, of vertices and edges [see for instance (Aingworth et al. 1996)]. Unfortunately, as soon as there are three criteria measured on infinite scales, this diameter has no upper bound, as expressed by the following theorem.

Theorem 2 (long explanations) *For any integer p, if there is a subset $A \subseteq N : |A| = 3$ and $\forall i \in A$, $|\mathbb{X}_i| \geq p$, then there is a relation $\mathcal{R}$ satisfying axioms 1, 2 and 3, and a pair $(x, y) \in \Delta_3$ such that $(x, y) \in \mathcal{E}_2(\mathcal{R})$ and any explanation of (x, y) by preference swaps of order at most 2 has a length greater than $2p$.*

Proof The proof requires instantiating the relation $\mathcal{R}$, and is presented in Appendix 1. We make use of the necessary preference relation introduced in the Sect. 3, for some carefully built preference information.   □

– *Values of the terms in the sequence* Another point concerns the choice of the values of the intermediate alternatives $e_1, \ldots, e_{n-1}$ on the different attributes. If these values are not chosen carefully, we believe they can induce a cognitive load to the decision maker, when she analyzes the sequence. Several options may be considered for these values. A "dynamic" option is to restrict the values of the attributes of $e_1, \ldots, e_{n-1}$ to the value of the attributes of x or y. This choice seems suitable to a decision context where there is only one statement $(x, y) \in \mathcal{R}$ to explain. However, the case may arise where the decision maker asks repeatedly for explanations for several statements, so that this policy would lead to intermediate alternatives having different values from one explained pair to the next. This issue may be solved considering a "static" option, where the values of the attributes $e_1, \ldots, e_{n-1}$ are restricted to a predefined list, independently of the pair (x, y), so that the intermediate alternatives always reference the same values on the attributes, hopefully reducing the workload for the decision maker. One option or the other may prove more or less convincing, depending on the context (see Sect. 4).

3 Necessary preference relation

3.1 Presentation of the relation

In many decision-aiding contexts, the preference relation $\mathcal{R}$ is not explicitly specified. It is often elicited: some amount of preference information is stated by the decision maker, which is extended by an algorithmic process. We use a holistic representation of the preference information, described as a finite collection $\mathcal{P} \subset \mathbb{X}^2$ of preference statements: $(x, y) \in \mathcal{P}$ stating that x is preferred to y.

---

[3] The diameter in the graph is the longest distance between two vertices in graph.

Example 3 (ex. 1, continued) The preference information elicited from the decision maker can be expressed by three preference statements. $\mathcal{P} := \{\pi_1, \pi_2, \pi_3\}$, with

$$\pi_1 := \big((4*, \text{no}, 15 \text{ min}, 180\$), \quad (2*, \text{yes}, 45 \text{ min}, 50\$)\big)$$
$$\pi_2 := \big((2*, \text{no}, 45 \text{ min}, 50\$), \quad (2*, \text{yes}, 15 \text{ min}, 180\$)\big)$$
$$\pi_3 := \big((2*, \text{yes}, 15 \text{ min}, 180\$), \quad (4*, \text{no}, 45 \text{ min}, 180\$)\big)$$

A model compatible with this preference information outputs a relation $\mathcal{R}_\mathcal{P} \supset \mathcal{P}$. For instance, a preference model can be built upon any value function $V \in \mathbb{R}^{\mathbb{X}}$ that assigns a value to each alternative, and gives precedence to the higher valued alternative.

Definition 7 (*value models*) $\forall V \in \mathbb{R}^{\mathbb{X}}$, $\mathcal{R}_V := \{(x, y) \in \mathbb{X}^2 : V(x) \geq V(y)\}$

Any value model is obviously transitive and satisfies Axiom 2 introduced in Sect. 2. To also satisfy Axioms 1 and 3, we require the value function to be separable.

Definition 8 (*additive value functions*) $\forall \mathcal{P} \subset \mathbb{X} \times \mathbb{X}$,

$$\mathbb{V} := \left\{ V \in \mathbb{R}^{\mathbb{X}} : V(x) = \sum_{i \in N} v_i(x_i) \text{ and } \forall i \in N, \ v_i \in \mathbb{R}^{\mathbb{X}_i} \text{ is non-decreasing} \right\}$$
$$\mathbb{V}_\mathcal{P} := \{V \in \mathbb{V} : \forall (x, y) \in \mathcal{P}, \ V(x) \geq V(y)\}$$

Proposition 1 (properties of additive value models) *(Krantz et al. 1971)* *For any value function $V \in \mathbb{V}$, the corresponding value model $\mathcal{R}_V$ satisfies Axioms 1, 2 and 3.*

Any additive value model can thus benefit from the explanation engine described in Sect. 2, as the conceits involved may prove difficult for a broad audience, especially when conclusions are drawn from the particular shape of the marginal value functions v_i.

The non-empty[4] set $\mathbb{V}_\mathcal{P}$ contains all the additive value functions compatible to $\mathcal{P}$, i.e., that correctly outputs each comparison in the preference information. While many decision frameworks, such as UTA, instantiate this model by specifying a single suitable function $V \in \mathbb{V}_\mathcal{P}$, the necessary preference relation (Greco et al. 2008) circumvents the arbitrary nature of the choice of a particular value function, by demanding that every value function compatible to $\mathcal{P}$ rates alternative x higher than alternative y to assess that x is necessarily preferred to y.

Definition 9 (*necessary preference relation inferred from $\mathcal{P}$*)

$$\forall \mathcal{P} \subset \mathbb{X} \times \mathbb{X}, \quad \mathcal{N}_\mathcal{P} := \{(x, y) \in \mathbb{X} \times \mathbb{X} : \forall V \in \mathbb{V}_\mathcal{P}, \quad V(x) \geq V(y)\}$$

---

[4] The set $\mathbb{V}_\mathcal{P}$ is not empty, as it contains at least all uniform value functions. It may sometimes come down to contain only these, if the preference information is somewhat inconsistent. Any uniform value function V_{uniform} leads to a degenerated, complete relation $\mathcal{R}_{V_{\text{uniform}}} \equiv \mathbb{X}^2$.

We believe this extra layer of abstraction added on top of the modelling of preference by additive value functions requires some supportive evidence, the more down to earth the better. Fortunately, the necessary preference relation $\mathcal{N}_\mathcal{P}$ qualifies for the explanation engine developed in Sect. 2, as it satisfies all three axioms made on the relation to be explained.

Theorem 3 : *The binary relation $\mathcal{N}_\mathcal{P}$ satisfies Axioms 1, 2 and 3*

Proof By definition, $\mathcal{N}_\mathcal{P} = \bigcap_{V \in \mathbb{V}_\mathcal{P}} \mathcal{R}_V$. By Theorem 1, every binary relation $\mathcal{R}_V$ satisfies Axiom 1 and is a superset of $\mathcal{D}$, and so is their intersection. Hence, $\mathcal{N}_\mathcal{P}$ satisfies Axiom 1.

Let (x, y) and (y, z) be two pairs in $\mathcal{N}_\mathcal{P}$. For any value function $V \in \mathbb{V}_\mathcal{P}$, both (x, y) and (y, z) are in $\mathcal{R}_V$ (by definition of the necessary preference relation), and the pair (x, z) is in $\mathcal{R}_V$ (by transitivity of $\mathcal{R}_V$, see Theorem 1). As $(x, z) \in \mathcal{R}_V$ for any $V \in \mathbb{V}_\mathcal{P}$, the pair (x, z) is in $\mathcal{N}_\mathcal{P}$, so $\mathcal{N}_\mathcal{P}$ is transitive and satisfies Axiom 2. It is straightforward to adapt this argument to prove $\mathcal{N}_\mathcal{P}$ also satisfies the cancelation axiom. □

In the remainder of this section, the preference information $\mathcal{P}$ is considered given once and for all, and we will omit the corresponding quantifier "$\forall \mathcal{P} \subset \mathbb{X}^2$".

3.2 The decision problem: basic principles

The inference of the relation $\mathcal{N}_\mathcal{P}$ from the preference information $\mathcal{P}$ amounts to solving many decision problems, queries of the form "is x necessarily preferred to y?", for every pair $(x, y) \in \mathbb{X}^2$.

This issue has already been addressed by various techniques.

- In the wake of the original article (Greco et al. 2008) introducing the relation $\mathcal{N}_\mathcal{P}$, decision over a query requires solving a linear program (LP) minimizing $V(x) - V(y)$ subject to constraints ensuring the additive value function V is compatible to both the preference information $\mathcal{P}$ and the Pareto dominance $\mathcal{D}$, then concluding that x is indeed preferred to y if and only if $\min V(x) - V(y)$ is non-negative.
- Trying to write rule-based conditions on so-called positive and negative arguments for necessary preference of x over y, as proposed by (Spliet and Tervonen 2014).

An issue sometimes mentioned [e.g., (Spliet and Tervonen 2014)] is that necessary preference is a tall order, often resulting to a quite small set $\mathcal{N}_\mathcal{P}$, so that most pairs $(x, y) \in \mathbb{X} \times \mathbb{X}$ end up being incomparable (that is, neither (x, y) nor (y, x) are in $\mathcal{N}_\mathcal{P}$). It should be noted though that $\mathcal{N}_\mathcal{P}$ is far from minimal:

- The transitive closure of $\mathcal{D} \cup \mathcal{P}$ does not generally satisfy Axiom 3, so it is usually a strict subset of $\mathcal{N}_\mathcal{P}$.
- $\mathcal{N}_\mathcal{P}$ is actually not minimal under Axioms 1, 2 and 3. Indeed, the necessary preference relation also satisfies an additional axiom of multiple cancelation, which will prove to be central in our setting.

To first illustrate the intuition behind this additional axiom, let us consider the following example:

Example 4 (example 3 continued) For any $V \in \mathbb{V}_\mathcal{P}$, the following inequalities stand:

– From $((4^*, \text{no}, 15 \text{ min}, 180 \$),(2^*, \text{yes}, 45 \text{ min}, 50 \$)) \in \mathcal{P}$ we derive:

$$u_*(4^*) + u_r(\text{no}) + u_t(15 \text{ min}) + u_\$(180\$) \geq u_*(2^*) + u_r(\text{yes})$$
$$+ u_t(45 \text{ min}) + u_\$(50\$)$$

– From $((2^*, \text{no}, 45 \text{ min}, 50 \$),(2^*, \text{yes}, 15 \text{ min}, 180 \$)) \in \mathcal{P}$ we derive:

$$u_*(2^*) + u_r(\text{no}) + u_t(45 \text{ min}) + u_\$(50\$) \geq u_*(2^*) + u_r(\text{yes})$$
$$+ u_t(15 \text{ min}) + u_\$(180\$)$$

– From dominance for the criterion restaurant we derive:

$$u_r(\text{yes}) \geq u_r(\text{no})$$

Adding these three inequalities, and canceling terms appearing on both sides leads to:

$$\forall V \in \mathbb{V}_\mathcal{P}, \ u_*(4^*) + u_r(\text{no}) \geq u_*(2^*) + u_r(\text{yes})$$

which in turn proves, for instance, the necessary preference of $(4^*, \text{no}, 15 \text{ min}, 50 \$)$ over $(2^*, \text{yes}, 15 \text{ min}, 50 \$)$.

Formally, this property is thus called multiple cancelation in the literature (Krantz et al. 1971; Fishburn 1997).[5] It has been established [see (Joel Michell 1988)] to be logically independent from the axiom of cancelation, and if $\mathbb{X}$ is large enough, there are relations in $\mathbb{X}^2$ that satisfy Axioms 1, 2 and 3, but not double cancelation.

Regarding our explanation objective, this principle is extremely attractive: it accounts for the inference of new pairs in $\mathcal{N}_\mathcal{P}$ by canceling arguments throughout multiple statements, as illustrated in the previous example, a feature that none of the other techniques offers. However, one can wonder if this situation, where a statement of $\mathcal{N}_\mathcal{P}$ is proven by combining a subset of the previously approved statements of $\mathcal{P}$ and $\mathcal{D}$, is the rule or a lucky exception. We now address this issue by introducing a new framework for the resolution of a query.

3.3 A novel technique to solve the decision problem

In this section, we present a decision framework for answering the query "is alternative x necessarily preferred to alternative y?", given a set of preference statements $\mathcal{P}$: if

---

[5] mth-order cancelation axiom: consider $m+1$ alternatives $x^{(k)}$ in $\mathbb{X}$, $k \in \{0, 1, \ldots, m\}$. Let $y^{(k)}$ in $\mathbb{X}$, $k \in \{0, 1, \ldots, m\}$ $m+1$ alternatives such that, for every criterion $i \in N$, $(y_i^{(0)}, y_i^{(1)}, \ldots, y_i^{(m)})$ is a permutation of $(x_i^{(0)}, x_i^{(1)}, \ldots, x_i^{(m)})$. Then, $[(x^{(k)}, y^{(k)}) \in \mathcal{R}, \forall k \in \{0, 1, \ldots, m-1\}] \Rightarrow (y^{(m)}, x^{(m)}) \in \mathcal{R}$.

(x, y) is an unbounded pair, as defined by Definition 13, then necessary preference does not hold (Theorem 4); else, we define covectors for the pair (x, y) (see Definition 12) permitting to express three characterizations of a positive query (Theorem 5): the absence of solution to a linear system of inequalities; the expression of the covector expressing the query as a linear combination with non-negative coefficients of the covectors of the preference statements and of the covectors of the dual base; a slightly modified version of this linear combination, where the coefficients sought for are non-negative integers.

The preference information references a finite set of attributes for each criterion. We call core alternatives the finite set of alternatives combining these attributes.

Definition 10 (*core alternatives*)

$$\mathbb{D}_i := \bigcup_{(x,y)\in\mathcal{P}} \{x_i, y_i\} := \{d_{i,1} \prec_i \cdots \prec_i d_{i,|\mathbb{D}_i|}\} \,;\, \mathbb{D} := \prod_{i\in N} \mathbb{D}_i$$

Example 5 (Example 3 continued)

$$\begin{aligned}
\mathbb{D}_* &= \{a \prec_* A\} &&\text{with } a := 2* \text{ and } A := 4* \\
\mathbb{D}_r &= \{b \prec_r B\} &&\text{with } b := \text{no and } B := \text{yes} \\
\mathbb{D}_t &= \{c \prec_t C\} &&\text{with } c := 45 \text{ min and } C := 15 \text{ min} \\
\mathbb{D}_\$ &= \{d \prec_\$ D\} &&\text{with } d := 180\$ \text{ and } D := 50\$
\end{aligned}$$

Consequently, the preference statements are: $\pi_1 = (AbCd, aBcD)$; $\pi_2 = (abcD, aBCd)$; $\pi_3 = (aBCd, Abcd)$ and there are 16 core alternatives: $\mathbb{D} = \{ABCD, ABCd, ABcD, ABcd, AbCD, AbCd, AbcD, Abcd, aBCD, aBCd, aBcD, aBcd, abCD, abCd, abcD, abcd\}$

In the remainder of this section, we often use interval semantics, where an interval designates all the attributes simultaneously higher than the lower bound and lower than the upper bound:

$$\forall i \in N, \forall a_i, b_i \in \mathbb{X}_i, [a_i, b_i] := \{z \in \mathbb{X}_i : a_i \precsim_i z \precsim_i b_i\}$$

In particular, core intervals $[d_{i,k}, d_{i,k+1}]$ play a key role. They are indexed by pairs (i, k) conveniently grouped in an index set $\mathbb{I}$:

Definition 11 (*indexes of core intervals*) The set $\mathbb{I} := \bigcup_{i\in N}\{(i, k) \,:\, k \in \mathbb{N} \text{ and } 1 \leq k \leq |\mathbb{D}_i| - 1\}$ contains the pairs (i, k) indexing the core intervals $[d_{i,k}, d_{i,k+1}]$ and, consequently, the differences in marginal value between consecutive core levels $\Delta v_{(i,k)} := v_i(d_{i,k+1}) - v_i(d_{i,k})$.

We denote $\times$ the matrix multiplication, so that, for a (line) covector $v^\star$ and a (column) vector w both taken in $\mathbb{R}^{\mathbb{I}}$, $v^\star \times w = \sum_{(i,k)\in\mathbb{I}} v^\star_{(i,k)} w_{(i,k)}$.

This collection of intervals $[d_{i,k}, d_{i,k+1}]$, $(i, k) \in \mathbb{I}$ is partitioned between pros, cons and neutral arguments of a pair of alternatives (x, y).
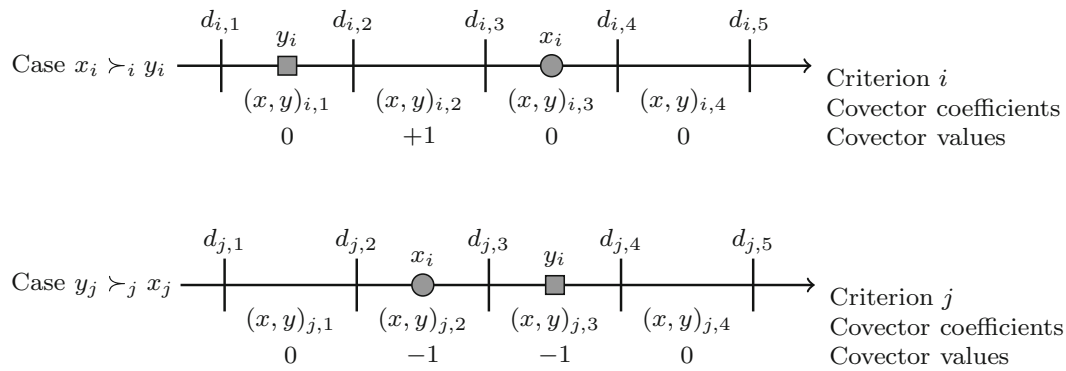
Fig. 1 Covectors illustrated

Definition 12 (*covector associated to a pair of alternatives*) $\forall (x, y) \in \mathbb{X}^2$, the covector $(x, y)^\star$ is a linear form operating on $\mathbb{R}^\mathbb{I}$. Its coefficient associated with criterion $i \in N$ and interval $[d_{i,k}, d_{i,k+1}] \subset \mathbb{X}_i$ is given by:

$$(x, y)^\star_{(i,k)} := \begin{cases} +1, & \text{if } [d_{i,k}, d_{i,k+1}] \subset [y_i, x_i] \\ -1, & \text{if } [d_{i,k}, d_{i,k+1}] \cap]x_i, y_i[\neq \emptyset \\ 0, & \text{else} \end{cases}$$

The canonical dual base is denoted $(\delta^\star_{(i,k)})_{(i,k) \in \mathbb{I}}$, where the covector $\delta^\star_{(i,k)}$ has all coefficients equal to zero, except for the coefficient associated to the interval indexed by (i, k), which is equal to +1, so that $\delta^\star_{(i,k)} \times \Delta v = \Delta v_{(i,k)}$.

For alternatives (x, y) in the core $\mathbb{D}^2$, for each criterion $i \in N$, intervals $[d_{i,k}, d_{i,k+1}]$ between x_i and y_i are taken into account, positively if $x_i \succ_i y_i$, and negatively if $y_i \succ_i x_i$. For alternatives (x, y) outside the core, for some criterion $i \in N$, some attribute x_i, or y_i, or both, falls strictly between the values of $\mathbb{D}_i$, "breaking" some interval $[d_{i,k}, d_{i,k+1}]$. Because of the cautious nature of the relation $\mathcal{N}_\mathcal{P}$, "broken" intervals are rounded down: those that would support the preference of x over y is not taken into account and considered neutral, with coefficient 0, while "broken" intervals that would go against this preference are totally taken into account with coefficient -1. Figure 1 illustrates these notions.

Example 6 As the preference information only refers two attributes level by criteria, there is exactly one core interval by criterion: from 2* to 4*, from no to yes, from 45 min to 15 min and from 180 \$ to 50 \$. Definition 12 is straightforward for core alternatives:

$$\begin{aligned} \pi_1 &= (AbCd, aBcD); & \pi_1^\star &= (1, -1, 1, -1); \\ \pi_2 &= (abcD, aBCd); & \pi_2^\star &= (0, -1, -1, 1); \\ \pi_3 &= (aBCd, Abcd); & \pi_3^\star &= (-1, 1, 1, 0). \end{aligned}$$

Alternatives outside the core demand a bit more effort: $(h_1, h_3)^\star = (0, 1, 0, -1)$, as:

– h_1 (5∗) is more comfortable than h_3 (3∗), but not strongly enough to warrant for a positive argument;
– h_1 is strongly better than h_3 on criterion restaurant;
– h_1 is weakly nearer than h_3;
– h_3 is weakly cheaper than h_1, and this counts as a fully negative argument.

We also find $(h_1,h_4)^\star = (1, 1, 1, -1)$, $(h_3, h_2)^\star = (-1, -1, 1, 1)$.

There is a class $\mathcal{U}_\mathcal{P}$ of unbounded queries (x, y) for which covectors fail to account for arguments that are both negative (because $y_i \succ_i x_i$) and infinitely strong (because $x_i \prec_i \min \mathbb{D}_i$ or $y_i \succ_i \max \mathbb{D}_i$). In such a case, x is clearly not necessarily preferred to y.

Definition 13 (*unbounded pairs $\mathcal{U}_\mathcal{P}$*)

$$\forall x, y \in \mathbb{X}, (x, y) \in \mathcal{U}_\mathcal{P} \iff \exists i \in N : x_i < y_i \text{ and } [x_i, y_i[\not\subseteq [\min \mathbb{D}_i, \max \mathbb{D}_i]$$

Theorem 4 : $\mathcal{U}_\mathcal{P} \cap \mathcal{N}_\mathcal{P} = \emptyset$

Proof : see Appendix 1. □

Example 7 (Example 5 continued) We see that h_3 is not necessarily preferred to h_1, as $(h_1)_* = 5*$ is better than both $(h_3)_* = 3*$ and the most comfortable hotel referenced by $\mathcal{P}$ (max $\mathbb{D}_* = 4*$). No amount of positive arguments in favor of h_3 make up for such a high attribute within the cautious context of necessary preference.

Neither is h_4 preferred to h_2, as $(h_4)_t = 60$ min is worse than both $(h_2)_t = 35$ min and the farthest hotel referenced by $\mathcal{P}$ (min $\mathbb{D}_t = 45$ min). No amount of arguments in favor of h_4 make up for such a low attribute.

For pairs outside the class $\mathcal{U}_\mathcal{P}$, we give three characterizations of the necessary preference of x over y using covectors.

Theorem 5 (characterization of necessary preference using covectors) $\forall(x, y) \in \mathbb{X}^2 \backslash \mathcal{U}_\mathcal{P}$, *the following propositions are equivalent:*

1. *Necessary preference*

$$(x, y) \in \mathcal{N}_\mathcal{P}$$

2. *Linear feasibility problem*

$$\begin{cases} & (x, y)^\star \times \Delta v < 0 \\ \forall \pi \in \mathcal{P}, & \pi^\star \times \Delta v \geq 0 \qquad \text{has no solution} \, \Delta v \in \mathbb{R}^\mathbb{I} \\ \forall (i, k) \in \mathbb{I}, & \delta^\star_{(i,k)} \times \Delta v \geq 0 \end{cases}$$

3. *Combination of statements* $\exists \lambda \in [0, +\infty[^\mathcal{P}, \ \mu \in [0, +\infty[^\mathbb{I}:$

$$(x, y)^\star = \sum_{\pi \in \mathcal{P}} \lambda_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} \mu_{(i,k)} \delta^\star_{(i,k)}$$

4. Integral combination of statements $\exists n \in \mathbb{N}^\star$, $\ell \in \mathbb{N}^\mathcal{P}$, $m \in \mathbb{N}^\mathbb{I}$:

$$n(x, y)^\star = \sum_{\pi \in \mathcal{P}} \ell_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} m_{(i,k)} \delta^\star_{(i,k)}$$

Proof: see Appendix 1. □

Point 3 proves the situation depicted in example 4 is not a corner case, but a general one: every necessary preference statement results from basic arithmetic operations (namely multiplication by a positive number, addition and cancelation of terms) over fundamental inequalities expressing either the preference information, or dominance. The exploration of the different combinations of this grammar, to assess if an alternative is necessarily preferred to another, is a linear programming problem. Noticeably, when the pair $(x, y) \notin \mathcal{U}_\mathcal{P}$ changes, the constraints remain the same, and can be computed once and for all: two different queries differ only by their objective covector.

Example 8 We use the fourth point of Theorem 5 to establish:

- h_1 is necessarily preferred to h_4, as $(h_1, h_4)^\star = \pi_1^\star + 2\delta^\star_{(2,1)}$;
- h_3 is necessarily preferred to h_2, as $(h_3, h_2)^\star = \pi_1^\star + 2\pi_2^\star + 2\pi_3^\star$;
- h_1 is not necessarily preferred to h_3, as there is no suitable linear combination.

Consequently, alternatives h_1 and h_3 are incomparable, as neither is preferred to the other.

We represent graphically the skeleton of the relation $\mathcal{N}_\mathcal{P} \cap \mathbb{D}^2$ (additional arcs resulting of the transitive closure of this skeleton are omitted in Fig. 2). For illustrative purpose, we show some example of the covectors associated to pairs involved in Example 4.

The integral version (point 4) is obviously less useful than the continuous one (point 3) for the actual decision of a query, as it implies the solving of an ILP, rather than an LP. It is nevertheless an important property that we shall leverage in the next section to derive insights into the problem of explaining a necessary preference relation statement $(x, y) \in \mathcal{N}_\mathcal{P}$ by low-order preference swaps, as introduced in Sect. 2.

4 Explanation of the necessary relation with binary reference scales

In this section, we bring together the main notions discussed in Sects. 2 and 3, connecting the explanation engine producing sequences of low-order preference swaps to the necessary preference relation. This coupling is made possible by Theorem 3, which ensures the necessary preference relation $\mathcal{N}_\mathcal{P}$ satisfies the requirement for the relation $\mathcal{R}$ explained by the explanation engine (i.e., we instantiate $\mathcal{R}$ as $\mathcal{N}_\mathcal{P}$). This coupling is also highly desirable, as the necessary preference relation makes minimal assumptions, handling a collection of compatible utility functions, virtually impossible to exhibit to the user.

To address some of the issues listed in Sect. 2.3, we make two additional assumptions. The first one concerns the number of distinct values referenced by the preference

Fig. 2 Necessary preference relations
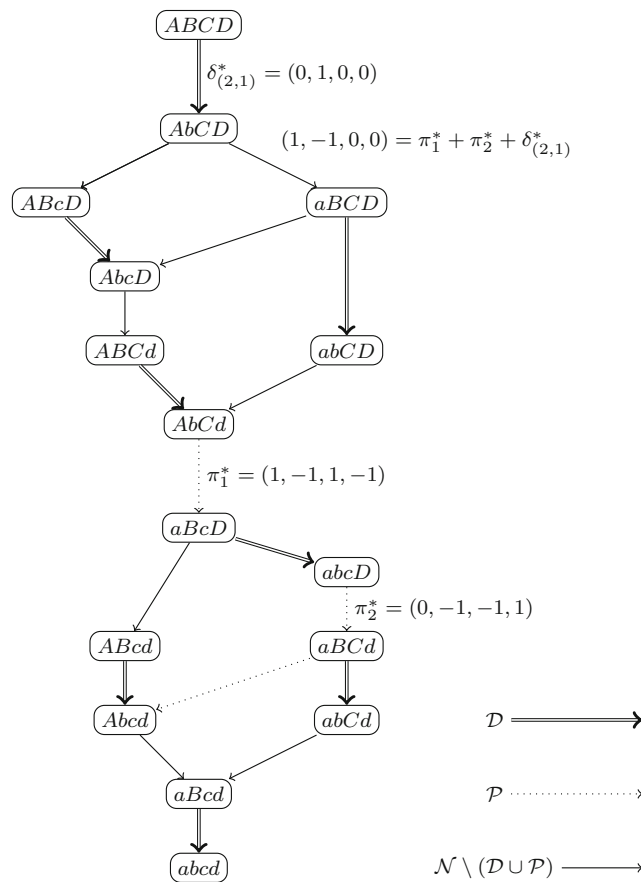
information $\mathcal{P}$ which serves as a basis for the inference of the necessary preference relation $\mathcal{N}_\mathcal{P}$, and is discussed in Sect. 4.1. The second one instantiates the cap on the order of the swaps linking the alternatives in the explaining sequence, and is discussed in Sect. 4.2. Under these assumptions, explanations have a core, term-by-term structure we expose in Sect. 4.3, followed by some resulting properties.

4.1 Binary reference scales

Binary reference scales are encountered when the preferences $\mathcal{P}$ expressed by the decision maker only reference two levels on each attribute.

Definition 14 (*Binary reference scales*)

$$\forall i \in N, \ \mathbb{B}_i = \{\top_i \succsim_i \bot_i\}, \ \mathbb{B} := \prod_{i \in N} \mathbb{B}_i$$

Besides luck, such a tight reference set is the consequence of one of these two situations :

– *Attributes are themselves binary*: present or absent features, passed or failed checks, etc. In addition, such binary attributes may result from any model relying

on subset comparisons. While they fall outside the scope of this article, we believe the explanation engine discussed here can address problems not necessarily resulting from an additive utility decision model (for instance, robust weighted majority decision models rely on subset comparisons between coalition of criteria, as do pan-balance comparisons encountered in extensive measurement problems).

– *When expressing preference statements, the decision maker is deliberately restricted to comparing between prototypical alternatives specifically chosen in* $\prod_{i \in N}\{\bot_i, \top_i\}$. This process is supposed to help the decision maker focusing on the main aspects of the preference problems, by limiting the number of changing parts between alternatives, and by referring to carefully chosen reference values, serving as anchors. This technique is used in the field of experimental design (yielding the one-factor-at-a-time or the factorial experiments methods), as well as in multicriteria decision aiding. For instance, the MACBETH method (Bana e Costa and Vansnick 1995; Bana e Costa et al. 2008) is based on binary alternatives: to assess hidden technical parameters (the weights of the various criteria), the decision maker is asked to express preference between prototypical alternatives, traditionally referencing a neutral level $\bot_i$ (for technological products, representing the attribute of a mid-range, available product), and a high-level $\top_i$ (representing the attribute of a luxury product, or a hypothetical performance demanding a technological breakthrough).

This tight set of core alternatives (see Definition 10) has bearing on the necessary preference relation. It increases the likelihood of single and multiple cancelation occurrence, thus enriching relation $\mathcal{N}_{\mathcal{P}}$ between core alternatives in $\mathbb{B}^2$. It aligns the individual technical arguments of the decision problem "is alternative x necessarily preferred to alternative y?", the intervals between consecutive attributes of the core (see Definition 12), with the criteria themselves. This alignment has, in turn, consequences concerning explanations, as the criteria involved in a preference statement (precisely, their number) determine its order, which is a proxy for its cognitive complexity. Technically, with binary reference scales, the order of a swap $(x, y) \in \mathcal{N}_{\mathcal{P}} \backslash \mathcal{D}$ is exactly the number of non-zero coefficients of its covector $(x, y)^\star$.

4.2 Swaps of order two

While the assumption of binary scales is a favorable case for the joining of the explanation engine based on sequences of preference swaps and the necessary preference relation, we make the choice concerning the bound placed on the order of the swaps eligible for participating in the explanation. We restrict the explanation to swaps of order at most two, that is:

– either a dominance relation or
– a trade-off between exactly two criteria.

The concept of swaps is known in engineering. For instance, the Architecture Trade-off Analysis Method (ATAM) is used to assess software architectures according to "quality attribute goals" (Kazman et al. 2000). A trade-off point is an architecture

parameter affecting at least two quality attributes in different directions. For example, increasing the speed of the communication channel improves throughput in the system but reduces its reliability. Thus, the speed of that channel is a trade-off point. The concept of trade-off point in ATAM makes explicit the interdependencies between attributes. Even though trade-offs can be defined for any number of attributes, the examples of trade-offs that are provided by experts are almost always given on pairs of attributes. This is the case of the example provided above. It is thus a very reasonable assumption to restrict ourselves to swaps of order two.

4.3 Structure of an explanation

Our restriction to binary scales allows us to introduce a simpler notation, in terms of positive or negative arguments:

Definition 15 (*pros and cons of a necessary preference statement*) If $\mathcal{P} \subset \mathbb{B}^2, \forall (x, y) \in \mathcal{N}_\mathcal{P}$,

$$(x, y)^+ := \{i \in N : (x, y)^\star_{(i,1)} = +1\} = \{i \in N : y_i \precsim_i \bot_i \prec_i \top_i \precsim_i x_i\}$$
$$(x, y)^- := \{i \in N : (x, y)^\star_{(i,1)} = -1\} = \{i \in N : \bot_i \precsim_i x_i \prec_i y_i \precsim_i \top_i\}$$

Assuming binary reference scales, the relation $\Delta_2 \subset \mathbb{X}^2$ between alternatives induces a relation between criteria $\tilde{\Delta}_2 \subset N^2$.

Definition 16 (*criteria swaps*) If $\mathcal{P} \subset \mathbb{B}^2$,

$$\tilde{\Delta}_2 := \{(i, i') \in N^2 : ((\top_i, \bot_{i'}), (\bot_i, \top_{i'}))_{cp} \subset \Delta_2\}$$

Note the use of the ceteris paribus syntax here (see Definition 1). We emphasize though that this relation is not suitable to being presented directly as an explanation. The reason is that it could be interpreted, sometimes erroneously, as giving more importance to criterion i than to criterion i'. While this interpretation seems practically correct in an elicitation framework similar to the MACBETH procedure (see Sect. 4.1), it is highly dependent of the values of $\mathbb{D}_i \times \mathbb{D}_{i'}$ referred by the preference information $\mathcal{P}$. To remain on the safe side, the relation $\tilde{\Delta}_2$ should only appear as a technical tool to produce an explanation.

Example 9 The necessary preference relation deduced from the preference information given in Example 3 contains the following compact criteria swap statements, represented in Fig. 3.

$$\tilde{\Delta}_2 = \{(*, r), (t, r), (\$, *), (\$, r), (\$, t)\}.$$

For instance, the compact criteria swap statement $(\$, r)$, represented by the arrow from $\$$ to r, means that an alternative ranking higher than D on attribute $\$$ and low on attribute r is necessarily preferred to one ranking low on $\$$ (between d and D) and high on r, attributes $*$ and t being equal: $\left((\_*, b, \_t, D), (\_*, B, \_t, d)\right)_{cp} = \{\left((x_*, b, x_t, D), (x_*, B, x_t, d)\right), \forall x_* \in \mathbb{X}_*, \forall x_t \in \mathbb{X}_t\} \subset \mathcal{N}_\mathcal{P}$.

Fig. 3 Binary relation between criteria

The following theorem reveals the core structure every explanation is built upon.

Theorem 6 (Term-by-term explanation) *If $\mathcal{P} \subset \mathbb{B}^2$, $\forall \sigma \in \mathcal{N}_\mathcal{P}$, the following propositions are equivalent:*

1. $\sigma \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$
2. $\exists a \in \mathbb{N}^\star, \gamma_1, \ldots, \gamma_q \in \Delta_2, \ell_1, \ldots, \ell_q \in \mathbb{N}, m_1, \ldots, m_n \in \mathbb{N}:$

$$a\sigma^\star = \sum_k \ell_k \gamma_k^\star + \sum_k m_k \delta_{(k,1)}^\star$$

3. *There is a matching of cardinality $|\sigma^-|$ in the graph of $\tilde{\Delta}_2 \cap (\sigma^+ \times \sigma^-)$.*
4. *There is an injection $\phi : \sigma^- \to \sigma^+$ such that $\forall k \in \sigma^-, (\phi(k), k) \in \tilde{\Delta}_2$.*

Proof See Appendix 1. □

In a nutshell, an explanation is a sequence where, at each step, a positive argument is used up to cancel an inferior negative argument and, eventually, every negative argument has been canceled. We highlight three consequences of this theorem:

– *If preferences only refer to swaps of order 2, then every necessary preference can be explained by swaps of order 2.* This is a potent existence result for explanations, and it provides a complete description of the necessary preference relation under the assumption of the decision maker expressing preferences between alternatives differing along two criteria only.

Corollary 1 (case of 2-order preference statements) *If $\mathcal{P} \subset \mathbb{B}^2$, and $\forall \pi \in \mathcal{P}, |N_\pi^{\neq}| = 2$ then $\mathcal{E}_2(\mathcal{N}_\mathcal{P}) = \mathcal{N}_\mathcal{P}$. i.e., for any statement $(x, y) \in \mathcal{N}_\mathcal{P}$, there exists an explanation of it in $\mathcal{E}_2(\mathcal{N}_\mathcal{P})$*

Proof By Theorem 1, $\mathcal{E}_2(\mathcal{N}_\mathcal{P}) \subset \mathcal{N}_\mathcal{P}$. Reciprocally, if $(x, y) \in \mathcal{N}_\mathcal{P}$, the implication 1. $\Rightarrow$ 4. of Theorem 5 ensures the existence of a linear combination with integral, non-negative coefficients $n(x, y)^\star = \sum_{\pi \in \mathcal{P}} \ell_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} m_{(i,k)} \delta_{(i,k)}^\star$. The assumption that $\forall \pi \in \mathcal{P}, |N_\pi^{\neq}| = 2$ entails $\mathcal{P} \subset \Delta_2$, so this linear combination satisfies proposition 2 of Theorem 6, thus $(x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$ by proposition 1.

– *Explanations can be kept short.* The next corollary proves that the size of the explanation is at most "half the number of criteria, rounded down, plus one", which appears manageable for the recipient of explanation.

Corollary 2 (short explanations) *If $\mathcal{P} \subset \mathbb{B}^2$, for any statement $(x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$, there exists an explanation with a length at most $\lfloor \frac{|N|}{2} \rfloor + 1$, where $\lfloor m \rfloor$ denotes the integer part of m.*

The bound $\lfloor \frac{|N|}{2} \rfloor + 1$ basically comes from the fact that $|(x, y)^-| \leq \lfloor \frac{|N|}{2} \rfloor$, which follows directly from item 4 of Theorem 6. The main asset of this theorem is that it is constructive. The explanation sequence will be provided in the next section.

Algorithm 1: FINDEXPLANATION

Data: a statement $\sigma = (x, y)$ to be explained, a set of preference statements $\mathcal{P}$.
Result: a matching of each negative argument by a stronger positive one.
1 Compute σ^+, σ^-
2 if $|\sigma^+| < |\sigma^-|$ **then**
3 $\quad$ return None
4 if $\sigma \notin \mathcal{N}_\mathcal{P}$ **then**
5 $\quad$ return None

6 Build the graph of $\tilde{\Delta}_2 \cap (\sigma^+ \times \sigma^-)$:
7 Initialize $\mathcal{G}$ as a graph with nodes $\sigma^+ \cup \sigma^-$ and no edge.
8 for $i \in \sigma^+$ **do**
9 $\quad$ **for** $j \in \sigma^-$ **do**
10 $\quad\quad$ **if** *the LP with $|N| + |\mathcal{P}|$ inequality constraints, $|N|$ equality constraints and $|N| + |\mathcal{P}|$ variables*
11 $\quad\quad\quad$ $\forall p \in \mathcal{P}, \ell_p \geq 0$
12 $\quad\quad\quad$ $\forall k \in N, m_k \geq 0$
13 $\quad\quad\quad$ $\forall k \in N, \sum_{p \in \mathcal{P}} \ell_p\, p_k^\star + m_k = 1$ *if $k = i$, -1 if $k = j$, 0 else.*
14 $\quad\quad$ *is feasible* **then**
15 $\quad\quad\quad$ add edge (i, j) to $\mathcal{G}$

16 Find a matching ϕ of maximum cardinality C in bipartite graph $\mathcal{G}$.
17 if $C < |\sigma^-|$ **then**
18 $\quad$ return None
19 return ϕ

- *Building an explanation, or ensuring there is none, is handled by an efficient algorithm (see Algorithm 1).* A quick inspection of the complexity reveals that in the first part of the algorithm, there are at most $\mathcal{O}(n^2)$ calls to a linear program (with n the number of criteria). This is followed by the resolution of a matching problem, which runs in its simpler version in $\mathcal{O}(n^3)$. Note that in theory, the number of constraints and variables of the LP may be exponential in n, because of the number of preference statements can be. In practice, this is of course highly unrealistic as it is too demanding for the decision maker. Finally, for a polynomially bounded number of preference queries, the algorithm is efficient.

Example 10 (Ex 8. ctd.) The pair (h_1, h_4) is in $\mathcal{N}_\mathcal{P}$. Its negative arguments are $(h_1, h_4)^- = \{*, r, \$\}$ and its positive arguments $(h_1, h_4)^+ = \{\$\}$. As there are more

Fig. 4 Matching returned by
Algorithm 1 with data of
Example 3

negative than positive arguments, the necessary preference of h_1 over h_4 cannot be explained by a sequence of preference swaps of order 1 or 2.

The pair (h_3, h_2) is also in $\mathcal{N}_\mathcal{P}$. $(h_3, h_2)^- = \{*, r\}$ and $(h_3, h_2)^+ = \{t, \$\}$.

Figure 4 shows the bipartite graph of the relation Δ_2 restricted to pairs of positive–negative arguments of the statement (h_3, h_2). The double arrows highlight a matching of cardinality 2, covering the negative arguments, as returned by Algorithm 1: $\{(\$, *), (t, r)\} \subset \tilde{\Delta}_2$. Therefore, the statement (h_3, h_2) can be explained by a sequence of preference swaps of order 2 and dominance relations.

To explain that $h_3 = (3*, \text{no}, 15 \text{ min}, 60 \text{ \$})$ is necessarily preferred to $h_2 = (4*, \text{yes}, 45 \text{ min}, 180 \text{ \$})$, several explanations can be considered:

- $h_3 \; \Delta_2 \; (4\star, \text{no}, 15 \text{ min}, 180\$) \; \Delta_2 \; h_2$
- $h_3 \; \Delta_2 \; (3\star, \text{yes}, 45 \text{ min}, 180\$) \; \Delta_2 \; h_2$
- $h_3 \; \mathcal{D} \; (a, b, C, D) \; \Delta_2 \; (A, b, C, d) \; \Delta_2 \; (A, B, c, d) \; \mathcal{D} \; h_2$
- $h_3 \; \mathcal{D} \; (a, b, C, D) \; \Delta_2 \; (a, B, c, D) \; \Delta_2 \; (A, B, c, d) \; \mathcal{D} \; h_2$

The first two explanations, which involve directly the attributes of the compared alternatives are shorter than the last two, which refer to core alternatives. It is interesting to observe how the two preference swaps (giving up cost for comfort and lengthening commute time to obtain access to a restaurant) can be presented in any order (since they do not have any criteria in common).

5 Related works and extensions

Generating explanations to justify recommendation is a key challenge to decision-aiding systems. While we witness the emergence of highly sophisticated methods to elicit preferences and compute recommended alternatives, the question of explanation is often neglected. We believe this may hinder the development of such systems. As a matter of fact, real decision makers often prefer the use of a very basic model if its outcomes are transparent, rather that elaborate models that look as a black box for them.

Explanations can either be conceived as being complete or incomplete. While we clearly follow the first option in this paper, some papers assume that explanations can be effective without being formally sufficient to support the statement [this may indeed be absolutely appropriate in settings with low stakes, for instance for most recommender systems (Herlocker et al. 2000; Friedrich and Zanker 2011)]. In that case, explanations can be seen as positive evidence supporting the conclusion. In a multicriteria setting close to ours, the approaches of Klein (1994), Carenini and Moore (2006), Labreuche (2011), Nunes et al. (2014) fall into that category: they build upon

patterns (or anchors) that are used to present some sufficiently convincing evidence to the user. The idea in that case is for instance to identify which set of criteria should be highlighted in the explanation.

A second distinctive feature of explanation is whether it is data based or process based (Herlocker et al. 2000). The vast majority of approaches dealing with this concept and emanating from A.I. adopts a data-based approach: this is true in particular of the literature investigating explanations in diagnosis systems [see for instance (Eiter and Gottlob 1995)] or constraints [where the aim is to return a minimal subset of mutually incoherent constraints in case of infeasibility (Ulrich Junker 2004)]. Here the objective is to find a minimal subset of the data provided by the user which implies the conclusion. This assumes that the explanation is to be presented to a user who has no problem in understanding the process by which these data then lead to a given conclusion. This is not the case in our setting (as inference from the necessary preference relation is a difficult notion to handle), and our approach follows instead a process-based approach. We would like to point out though that these two approaches are by no means contradictory: in particular, it would be certainly relevant to incorporate some data-based consideration when building sequences of preference swaps, as was already alluded to in the paper. Giving priority to the statements presented by the user, or defining notions of proximity so that sequences of explanations can be evaluated with respect to their distance to the initial data is certainly a promising perspective.

In our setting the initial preference information is provided as comparisons between alternatives. Other form of input may justify the use of other decision models (and consequently, of explanation techniques). For instance, complete explanations have been investigated for (weighted) majority-based decision models, when ordinal rankings on alternatives are given as input (Labreuche et al. 2011, 2012). In that case, explanations also amounts to exhibit coalitions of criteria.

Each explanatory step produced by our approach is typically performed by focusing on trade-offs on a subset of criteria, assuming the other ones remain unchanged. This ceteris paribus principle, which lies at the heart of the initial even-swap technique, has also been exploited for its ability to compactly represent qualitative conditional preferences (Boutilier et al. 2004). This language was later extended to account for possible trade-offs among criteria (Brafman et al. 2006), and (Nic Wilson 2011) proposed an even more expressive language (allowing to capture also stronger semantics). The resulting statements are similar in spirit to the criteria swaps that we use in this paper as technical constructs. Interestingly, "flipping" or swapping sequences appear as proof-theoretical counterpart for the semantics of these logical theories. While such compact statements are certainly useful for users to express preferences, it is not clear whether they should be used per se in producing explanations, because they may be inappropriately interpreted, as discussed in Sect. 4.3. Investigating their relevance in our setting is nevertheless an interesting future work.

We conclude by mentioning some further perspectives of this work.

– There remain theoretical questions to be studied. We have investigated two extreme cases: in the first one, no assumption is made on the preference information (yielding a negative result in terms of the length of the explanation), while in the second one we assume a binary reference scale (and can guarantee the existence of a short

explanation). A natural but challenging question is whether the complexity of the reference scale can be more generally linked to the size of the explanations.

– We have provided an algorithm for the binary case only. It would be of practical interest to design and implement an algorithm finding the simplest (e.g., shortest) explanation in the general case.

– While we discuss good theoretical properties of explanations, an empirical validation remains to be conducted on other aspects mentioned (the sequencing of swaps, the choice of values, for instance). What makes the exercise difficult though is that this may highly depend on the context of use: a DM who needs to justify an important decision before a committee may not have the same expectations as a DM taking a decision for herself. Other issues are likely to emerge too: in particular, as we saw in Example 10, the same preference swaps can (sometimes) be presented in different orders. Are there good heuristics to select a given ordering?

– The framework may be smoothly extended to cater for more general situations. For instance, the nature of the preferential information may be different. The DM may use a more expressive language, and give some statements on the intensity of their preferences. A first step in that direction is to assume a quaternary relation, of the form "o_1 is more intensely preferred to o_2 than o_3 is preferred to o_4". While this would constitute a first step towards dealing with intensities, we are confident that this may still be handled within the framework described here.

– As a final suggestion on a possible extension of this framework, we note that this work makes the assumption that elicitation and explanation are dealt with separately. A certainly promising perspective is to extend the framework so that explanation and elicitation are actually intertwined. By putting forward an explanation, the system shows some evidence which can in turn trigger some reaction from the DM.

Proofs

Proof of theorem 2

For the sketch of the proof, we construct, for every p, a preference between $x = (0, 0, 0)$ and $y = (2p, -p, -p)$. Starting from alternative $(0, 0, 0)$, we begin with a preference swap between attributes 1 and 2 (adding value 1 on the first attribute, and subtracting 1 on the second one). Then we perform a preference swap between attributes 1 and 3 (adding value 1 on the first attribute, and subtracting 1 on the third one). We proceed then again by a preference swap between attributes 1 and 2, and so on (the sequence is depicted in Fig. 5).

Proof (Theorem 2) The proof is based on an instantiation of $\mathcal{R}$ with the necessary preference relation. This latter is inferred from information $\mathcal{P}$, and is denoted by $\mathcal{N}_{\mathcal{P}}$. Let $n = 3$, $p \in \mathbb{N}^*$. Assume that $\mathbb{X}_1 \supseteq \{0, 1, 2, \ldots, 2p\}$, $\mathbb{X}_2 \supseteq \{-p, -p + 1, \ldots, -1, 0\}$ and $\mathbb{X}_3 \supseteq \{-p, -p+1, \ldots, -1, 0\}$. Consider the following preference information $\mathcal{P}$:

Fig. 5 Description of the sequence

$$\forall j \in \{0, \ldots, p-1\}$$
$$(((2j)_1, (-j)_2), ((2j+1)_1, (-j-1)_2))_{cp} \subset \mathcal{P} \qquad (1)$$
$$\forall j \in \{0, \ldots, p-1\}$$
$$(((2j+1)_1, (-j)_3), ((2j+2)_1, (-j-1)_3))_{cp} \subset \mathcal{P} \qquad (2)$$

where (1) [resp. (2)] correspond to a ceteris paribus pair on attributes $\{1, 2\}$ (resp. $\{1, 3\}$). Hence, $\mathbb{D}_1 = \{0, 1, 2, \ldots, 2p\}$, $\mathbb{D}_2 = \{-p, -p+1, \ldots, -1, 0\}$ and $\mathbb{D}_3 = \{-p, -p+1, \ldots, -1, 0\}$.

We set $x = (0, 0, 0)$ and $y = (2p, -p, -p)$. With this $\mathcal{P}$, we clearly obtain the sequence

$$(x, (1, -1, 0)) \in \mathcal{P} \qquad \text{(by (1))}$$
$$((1, -1, 0), (2, -1, -1)) \in \mathcal{P}, \ldots \qquad \text{(by (2))}$$
$$((2p-2, -(p-1), -(p-1)), (2p-1, -p, -(p-1))) \in \mathcal{P} \qquad \text{(by (1))}$$
$$((2p-1, -p, -(p-1)), (2p, -p, -p)) \in \mathcal{P} \qquad \text{(by (2))}$$

so that $(x, y) \in \mathcal{R}$. This sequence is of length $2p$.

There remains to prove that this is the shortest explanation.

To this end, we first need to determine the form of Δ_2. By Theorem 4, the necessary preference relation cannot hold outside the interval between the minimal and maximal elements of $\mathbb{D}$. Moreover, according to Theorem 5, the necessary preference relation between two alternatives z, z' holds iff a linear problem involving the covector of (z, z') is feasible. From these results, checking whether $(z, z') \in \mathcal{N}_\mathcal{P}$ is equivalent to checking boundness on z and z', and also checking whether $(t, t') \in \mathcal{N}_\mathcal{P}$ where $t, t' \in \mathbb{D}$ are appropriately chosen from z and z'. Therefore, we need only to consider the elements in Δ_2 that belong to $\mathbb{D}_1 \times \mathbb{D}_2 \times \mathbb{D}_3$ (The other ones can be deduced by Pareto dominance). The preference information (1) and (2) is very specific. In

particular, any value $k \in \mathbb{D}_1$ appears only in two examples—one in which k appears in the left-hand side [in (1)] and the other one where k appears in the right-hand side [in (2)]. Moreover, we notice that, in (1) and (2), the value on the first attribute is always increasing from the left-hand side to the right-hand side, and the value of the second and the third attributes is decreasing from the left-hand side to the right-hand side. Hence, the elements of Δ_2 cannot be obtained by a combination of two or more preference information. They are obtained only from one preference information [(1), (2)] and Pareto dominance $\mathcal{D}$. More precisely, Δ_2 is composed of the following pairs

$$\Big((i, j, k), \ (i', j', k') \Big)$$

where either there exists l such that $i = 2l$, $j = 2l + 1$, $j \geq -l > -l - 1 \geq j'$ and $k = k'$, or there exists l such that $i = 2l + 1$, $j = 2l + 2$, $j = j'$ and $k \geq -l > -l - 1 \geq k'$. From this, one can readily see that the explanation of the preference of x over y described earlier is the shortest one. $\qquad \square$

Proof of Theorem 4 and Theorem 5

Proof of (1) $\Longleftrightarrow$ (2)

The belonging of a pair of alternatives to the necessary preference relation can be expressed as a mathematical program. We have to prove that when the pair is not unbounded, its constrains and objective function are linear and can be expressed using the proposed, fixed-length covectors.

Pairs of core alternatives, and in particular, preference statements, are never unbounded. We begin by introducing $\forall (i, k) \in \mathbb{I}$, $\Delta v_{i,k} := v_i(d_{(i,k+1)}) - v_i(d_{(i,k)})$ and proving covectors, when applied to such a vector Δ_v of differences in value, correctly compute the difference of value between core alternatives.

We break down the Definition 12 by criterion:
$\forall i \in N, \quad \forall x_i, y_i \in \mathbb{X}_i, \quad \text{let } (x_i, y_i) \in \mathbb{R}^{|\mathbb{D}_i|-1} : \forall k \in \mathbb{N} : 1 \leq k \leq |\mathbb{D}_i| - 1,$

$$(x_i, y_i)_k^\star := \begin{cases} +1, & \text{if } [d_{i,k}, d_{i,k+1}] \subset [y_i, x_i] \\ -1, & \text{if } [d_{i,k}, d_{i,k+1}] \cap]x_i, y_i[\neq \emptyset \\ 0, & \text{else} \end{cases}$$

So that $\forall x, y \in \mathbb{X}, \forall (i, k) \in \mathbb{I}, \ (x, y)_{(i,k)}^\star = (x_i, y_i)_k^\star$.

Lemma 1 (expression of differences in value as a product)

$$\forall i \in N, \forall x_i, y_i \in \mathbb{D}_i, \forall V \in \mathbb{V}, \quad v_i(x_i) - v_i(y_i) = \sum_{k=1}^{|\mathbb{D}_i|-1} (x_i, y_i)_k^\star \Delta v_{(i,k)}$$

Proof First, we note that for any valid indexes $k_1 < k_2$, $\sum_{k=k_1}^{k_2} \Delta v_{(i,k)} = v_i(d_{i,k_2}) - v_i(d_{i,k_1})$

Second, we detail $\sum_{k=1}^{|\mathbb{D}_i|-1}(x_i, y_i)_k^\star \Delta v_{(i,k)}$, according to the sign of $x_i - y_i$:

– If $x_i > y_i$, the interval $]x_i, y_i[$ is empty, so the case leading to a coefficient $(x, y)_{(i,k)}^\star = -1$ does not occur. Non-zero coefficients correspond to intervals $[d_{i,k}, k_{i,k+1}[$ partitioning $[y_i, x_i[$, so that $\sum_{k=1}^{|\mathbb{D}_i|-1}(x_i, y_i)_k^\star \Delta v_{(i,k)} = (+1)(v_i(x_i) - v_i(y_i))$

– If $x_i < y_i$, the interval $[y_i, x_i]$ is empty, so the case leading to a coefficient $(x, y)_{(i,k)}^\star = +1$ does not occur. Non-zero coefficients correspond to intervals $[d_{i,k}, k_{i,k+1}[$ partitioning $[x_i, y_i[$, so that $\sum_{k=1}^{|\mathbb{D}_i|-1}(x_i, y_i)_k^\star \Delta v_{(i,k)} = (-1)(v_i(y_i) - v_i(x_i)) = v_i(x_i) - v_i(y_i)$

– If $x_i = y_i$, the interval $[x_i, y_i]$ is trivial and the interval $]x_i, y_i[$ is empty, so every coefficient $(x, y)_{(i,k)}^\star$ is equal to zero. Consequently, $\sum_{k=1}^{|\mathbb{D}_i|-1}(x, y)_{(i,k)}^\star \Delta v_{(i,k)} = 0 = v_i(x_i) - v_i(y_i)$.

Thus, $\forall i \in N, \ v_i(x) - v_i(y) = \sum_{k=1}^{|\mathbb{D}_i|-1}(x, y)_{(i,k)}^\star \Delta v_{(i,k)}$.   □

For any alternatives $x, y \in \mathbb{D}$, summing up these equalities over every criteria yields $V(x) - V(y) = (x, y)^\star \times \Delta_v$

Introducing $\forall x, y \in \mathbb{X}, \ \Delta V_{\inf}(x, y) := \inf_{V \in \mathbb{V}_{\mathcal{P}}} V(x) - V(y) \in \mathbb{R} \cup \{-\infty\}$, Definition 9 states that

$$\forall x, y \in \mathbb{X}, \ (x, y) \in \mathcal{N}_{\mathcal{P}} \iff \Delta V_{\inf}(x, y) \geq 0$$

In the case of pairs of core alternatives, the objective function as well as the constraints of the minimization problem $\Delta V_{\inf}(x, y)$ can be expressed using covectors and matrix multiplication, as permitted by Lemma 1, so that $\Delta V_{\inf}(x, y)$ is a linear program.

Lemma 2 (query between core alternatives)

$$\forall x, y \in \mathbb{D}, \ \Delta V_{\inf}(x, y) = \inf (x, y)^\star \times \Delta v \ \text{s.t.} \ \Delta v \in \Omega_{\mathcal{P}} \cap \Omega_{\mathcal{D}}$$

with $\Omega_{\mathcal{P}} := \{\Delta v \in \mathbb{R}^{\mathbb{I}} : \forall \pi \in \mathcal{P}, \ \pi^\star \times \Delta v \geq 0\}$ *and* $\Omega_{\mathcal{D}} := \{\Delta v \in \mathbb{R}^{\mathbb{I}} : \forall (i, k) \in \mathbb{I}, \ \delta_{(i,k)}^\star \times \Delta v \geq 0\}$.

Generally, with alternatives (x, y) not necessarily belonging to the core $\mathbb{D}$, it has been shown Greco et al. (2008) that minimizing $V(x) - V(y)$ over $V \in \mathbb{V}_{\mathcal{P}}$ is still a linear program, with additional decision variables accounting for the distinct values $\{x_i, y_i\} \notin \mathbb{D}_i$. The $v_i(x_i), v_i(y_i)$ are only constrained by the monotonicity of the marginal value functions, so the problem is separate:

$$\Delta V_{\inf} = \inf_{\Delta v \in \Omega_{\mathcal{P}} \cap \Omega_{\mathcal{D}}} \sum_{i \in N} \inf_{\substack{v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i}} v_i(x_i) - v_i(y_i)$$

with, $\forall i \in N,$
$$\begin{cases} UX_i := \{v_i(x_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{y_i\}, \quad z_i \succsim_i x_i \Rightarrow v_i(z_i) \geq v_i(x_i)\} \\ LX_i := \{v_i(x_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{y_i\}, \quad z_i \precsim_i x_i \Rightarrow v_i(z_i) \leq v_i(x_i)\} \\ UY_i := \{v_i(y_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{x_i\}, \quad z_i \succsim_i y_i \Rightarrow v_i(z_i) \geq v_i(y_i)\} \\ LY_i := \{v_i(y_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{x_i\}, \quad z_i \precsim_i y_i \Rightarrow v_i(z_i) \leq v_i(l_i)\} \end{cases}$$

Thus, it is possible to circumvent this augmentation of the decision space by:

– Considering a given criterion $i \in N$ and a given vector $\Delta v \in \Omega_{\mathcal{P}} \cap \Omega_{\mathcal{D}}$;
– Directly assigning the additional decision variables to their optimal values in the inner linear program

$$\inf_{v_i(x_i), v_i(y_i)} v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases};$$

– Checking this optimal case is correctly represented, either by an unbounded pair or in covector form.

We begin by focusing on the case where the values of $\mathbb{D}_i \cup \{x_i, y_i\}$ are all different. We sort these values in strictly ascending order, and we detail three cases according to the position of x_i and y_i amongst these $|\mathbb{D}_i| + 2$ values:

– The interval $[x_i, y_i]$ overflows the set $\mathbb{D}_i$, so that the pair $(x, y) \in \mathcal{U}_{\mathcal{P}}$ is unbounded. This case actually encompasses three subcases
– x_i has no predecessor, when x_i is the least element of $\mathbb{D}_i \cup \{x_i, y_i\}$. There is no constraints in $LX_i = \mathbb{R}$;
– y_i has no successor, when y_i is the highest element of $\mathbb{D}_i \cup \{x_i, y_i\}$. There are no constraints in $UY_i = \mathbb{R}$;
– Both preceding cases are simultaneously satisfied.

In any case,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = -\infty,$$

thus $V_{\inf}(x, y) = -\infty$ and $(x, y) \notin \mathcal{N}_{\mathcal{P}}$, thus proving Theorem 4;

– y_i is the predecessor of x_i, *so* x_i is the successor of y_i. In this case, the constraints UX_i, LX_i, UY_i, LY_i can all be replaced by the single equality $v_i(x_i) = v_i(y_i)$, which defines a solution both feasible and where the objective function is minimized with respect to the decision variables $v_i(x_i), v_i(y_i)$. Meanwhile, we consider the coefficients $(x, y)^{\star}_{(i,k)}$, $1 \le k < |\mathbb{D}_i|$: the interval $[y_i, x_i]$ does not contain a single core value $d_{i,k} \in \mathbb{D}_i$, hence $(x, y)^{\star}_{(i,k)} \ne +1$; the interval $]x_i, y_i[$ is empty, hence $(x, y)^{\star}_{(i,k)} \ne -1$; finally $(x, y)^{\star}_{(i,k)} = 0$. This proves the identity:

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = \sum_{k=1}^{|\mathbb{D}_i|-1} (x, y)^{\star}_{(i,k)} \Delta u_{(i,k)},$$

as both sides are equal to zero.

– x_i has a predecessor *which is not* y_i, and y_i has a successor which is not x_i. First, we rewrite $\inf v_i(x_i) - v_i(y_i)$ s.t. $\begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases}$ as a difference in marginal value between surrogate alternatives in the core $\mathbb{D}_i$. The predecessor $\underline{x_i}$ of x_i is given by $\underline{x_i} := \max\{d \in \mathbb{D}_i, d \precsim_i x_i\}$, so that the constraints UX_i, LX_i can

both be replaced by the single equality $v_i(x_i) = v_i(\underline{x_i})$, which defines a solution both feasible and where $v_i(\underline{x_i})$ is minimal with respect to the decision variable $v_i(x_i)$. The successor $\overline{y_i}$ of y_i is given by $\overline{y_i} := \min\{d \in \mathbb{D}_i, d \succsim_i y_i\}$, so that the constraints UY_i, LY_i can both be replaced by the single equality $v_i(y_i) = v_i(\overline{y_i})$, which defines a solution both feasible and where $v_i(y_i)$ is maximal, so the objective function is minimal, with respect to the decision variable $v_i(y_i)$.

Thus,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = v_i(\underline{x_i}) - v_i(\overline{y_i})$$

Second, as both surrogate alternatives $\underline{x_i}, \overline{y_i}$ belong to $\mathbb{D}_i$, Lemma 1 ensures that

$$v_i(\underline{x_i}) - v_i(\overline{y_i}) = \sum_{k=1}^{|\mathbb{D}_i|-1} (\underline{x_i}, \overline{y_i})_k^\star \Delta u_{(i,k)}$$

Third, we check that the covector coefficients for criterion i of the original pair match those of the surrogate pair, that is:

$$\forall k \in \mathbb{N} : 1 \le k < |\mathbb{D}_i|, \ (x_i, y_i)_k^\star = (\underline{x_i}, \overline{y_i})_k^\star$$

The proof is straightforward:

- If $x_i \succ_i y_i$, then there is at least one attribute value $d \in \mathbb{D}_i$ between x_i and y_i, so that the predecessor of x_i and the successor of y_i are in the same order, thus $\underline{x_i} \succsim_i \overline{y_i}$. Hence, the coefficient indexed by (i, k) of their respective covectors are in $\{0, +1\}$, with value $+1$, respectively, when $y_i \precsim_i d_{i,k} \prec_i d_{i,k+1} \precsim_i x_i$ and when $\overline{y_i} \precsim_i d_{i,k} \prec_i d_{i,k+1} \precsim_i \underline{x_i}$. The definition of the surrogate pair ensures these conditions are equivalent.
- If $x_i \prec_i y_i$, then obviously $\underline{x_i} \precsim_i \overline{y_i}$. Hence, the coefficients of their respective covectors indexed by (i, k) are in $\{0, -1\}$, with value 0, respectively, when $y_i \precsim_i d_{i,k}$ or $d_{i,k+1} \precsim_i x_i$, and when $\overline{y_i} \precsim_i d_{i,k}$ or $d_{i,k+1} \precsim_i \underline{x_i}$. The definition of the surrogate pair ensures these conditions are equivalent. Thus,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = \sum_{k=1}^{|\mathbb{D}_i|-1} (x_i, y_i)_k^\star \Delta u_{(i,k)}$$

The cases where $|\mathbb{D}_i \cup \{x_i, y_i\}| = |\mathbb{D}_i| + 1$ are correctly handled in the discussion above: if overflow (when either $x_i \prec_i \min \mathbb{D}_i$ or $y_i \succ_i \max \mathbb{D}_i$) does not occur, the case $x_i = y_i$ extends the case where the optimal value of $v_i(x_i) - v_i(y_i)$ is zero; the case where $y_i \in \mathbb{D}_i$ leads to the introduction of $\overline{y_i} := y_i$, and the case where $x_i \in \mathbb{D}_i$ leads to $\underline{x_i} := x_i$.

Finally, for any pair $(x, y) \in \mathbb{X}^2$, we have proven that, in every case, either the pair is unbounded and not in the relation $\mathcal{N}_{\mathcal{P}}$, or it can be represented by a covector such that $\Delta V_{\inf}(x, y) = \inf_{\Delta v \in \mathbb{R}^{\mathbb{I}}} (x, y)^{\star} \times \Delta v$ s.t. $\begin{cases} \forall \pi \in \mathcal{P}, \ \pi^{\star} \times \Delta v \geq 0 \\ \forall (i, k) \in \mathbb{I}, \ \delta^{\star}_{(i,k)} \times \Delta v \geq 0 \end{cases}$

Proof of (2) $\iff$ (3)

By Farkas' lemma, the problem (2) has no solution if, and only if, the objective linear form $(x, y)^{\star}$ is a linear combination with non-negative coefficients of the constraint linear forms $\{\pi^{\star}, \pi \in \mathcal{P}\}$ and $\{\delta^{\star}_{i,k}, (i, k) \in \mathbb{I}\}$.

Proof of (3) $\iff$ (4)

Obviously, (4) $\Rightarrow$ (3). Conversely, as the covectors involved in (3) have integral coordinates, the non-negative coefficients $\{\lambda_{\pi}, \pi \in \mathcal{P}\}$ and $\{\mu_{(i,k)}, (i, k) \in \mathbb{I}\}$, if they exist, can be chosen in the field of rational numbers. Multiplying the relation by the common denominator $n \in \mathbb{N}^{\star}$ of these coefficients leads to (4).

Proof of Theorem 6

We prove Theorem 6 in four steps: (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (1).

- (1) $\Rightarrow$ (2): Assume a statement $\sigma := (x, y) \in \mathcal{E}_2(\mathcal{N}_{\mathcal{P}})$. By Theorem 1 and Definition 5, there is an integer n and a tuple $(e_0, e_1, \ldots, e_n) \in \mathbb{X}^n$ such that $e_0 = x, e_n = y$ and $(e_j, e_{j+1}) \in \mathcal{D} \cup \Delta_2$ for any integer $j < n$. This transitive chain of dominance relations and swaps of order 2 can be transformed into the covector relation sought, by induction on the length of the explanation, as described by the following lemmas:

Lemma 3 (covector representation of dominance relations)

$$\forall \rho \in \mathcal{D}, \exists q \in \{0, +1\}^I : \rho^{\star} = \sum_{(i,k) \in \mathbb{I}} q_{(i,k)} \delta^{\star}_{(i,k)}$$

Proof A dominance relation has no negative argument, so its covector coefficient, given by Definition 12, is in $\{0, +1\}$. $\qquad \square$

Lemma 4 (covector representation of transitivity relations)

$$\forall x, y, z \in \mathbb{X}, \exists q \in \mathbb{N}^{\mathbb{I}} : (x, z)^{\star} = (x, y)^{\star} + (y, z)^{\star} + \sum_{(i,k) \in \mathbb{I}} q_{(i,k)} \delta^{\star}_{(i,k)}$$

Proof For core alternatives $x, y, z \in \mathbb{D}$, for any separate value function $V \in \mathbb{V}$,

$$(x, z)^{\star} \times \Delta v = V(x) - V(z)$$
$$= (V(x) - V(y)) + (V(y) - V(z))$$

$$= (x, y)^\star \times \Delta v + (y, z)^\star \times \Delta v$$
$$= ((x, y)^\star + (y, z)^\star) \times \Delta v$$

As the relation above stands for any vector $\Delta v \in [0, +\infty[$, it yields $(x, z)^\star = (x, y)^\star + (y, z)^\star = (x, y)^\star + (y, z)^\star + \sum_{(i,k)\in\mathbb{I}} q_{(i,k)}\delta^\star_{(i,k)}$ with $q = 0$.

For alternatives not necessarily in the core, and for any criterion $i \in N$, the trivial cases where $y_i \in \{x_i, z_i\}$, the case where $x_i = z_i$, or the case where x_i, y_i, z_i are all distinct, divided into 6 subcases considering the order of attributes x_i, y_i, z_i, all lead to $(x, z)^\star \geq (x, y)^\star + (y, z)^\star$ because of the rounding down of broken intervals occurring once in the LHS and twice in the RHS. As both sides are covectors with integer coefficients, the difference $(x, z)^\star - ((x, y)^\star + (y, z)^\star)$ is a covector with non-negative integer coefficients $q_{(i,k)}$. □

- (2) $\Rightarrow$ (3): Suppose there exists integer coefficients $a, \ell_1, \ldots, \ell_q, m_1, \ldots, m_n$ and preference swaps of order 2: $\gamma_1, \ldots, \gamma_q$ such that

$$a\sigma^\star = \sum_k \ell_k \gamma_k^\star + \sum_k m_k \delta^\star_{(k,1)} \tag{3}$$

Multiplying both sides of the covector Equation (3) by the vector $(1, \ldots, 1)$, we obtain the relation:

$$M := a(|\sigma^+| - |\sigma^-|) = \sum m_k \geq 0$$

To homogenize the right-hand side, we represent the dominance relation thanks to a dummy criterion: $N' = N \cup \{0\}$ so that $\tilde{\Delta}_1 := \{(i, 0), i \in N\} \subset N'^2$. Thus, relation $\mathcal{D} \cup \Delta_2$ is a graph with nodes in N'. Re-indexing coefficients ℓ_k by the positive and negative arguments of swap γ_k (summing up duplicates if needed), and introducing $\ell_{k,0} := m_k$:

$$a\,\sigma^\star = \sum_{\gamma\in\tilde{\Delta}_1\cup\tilde{\Delta}_2} \ell_{\gamma^+,\gamma^-}\gamma^\star \tag{4}$$

To complete the flow ℓ, we introduce:

- A source s supplying flow $\ell_{s,i} = a$ to the positive arguments $i \in \sigma^+$;
- A sink t collecting flow $\ell_{j,t} = a$ from the negative arguments $j \in \sigma^-$, and $\ell_{0,t} = M$ from node 0.

Covector Equation (4) ensures ℓ defines a feasible flow on the graph $(N' \cup \{s, t\}, \tilde{\Delta}_1 \cup \tilde{\Delta}_2 \cup \{s\} \times \sigma^+ \cup \sigma^- \times \{t\} \cup \{(0, t)\})$, without capacity constraints, as projection on the ith coordinate ensures flow conservation for node $i \in N$. Flow ℓ can be decomposed as a superposition of:

- Cycles, involving necessary equivalence between the nodes, and not contributing to the value of the flow;

 – Paths from the source s to the sink t passing through node 0, denoting a dominance relation. Their total contribution to the value of the flow is M;
 – Paths from the source s to the sink t not passing through node 0, with an overall contribution of $a \times |\sigma^-|$ to the value of the flow. Each of these paths links a positive argument $i_1 \in \sigma^+$ to a negative argument $i_r \in \sigma^-$ through necessary preference swaps of order 2. Transitivity of the necessary preference relation entails that i_1 is necessarily preferred to i_r: the edge (i_1, i_r) belongs to $\Delta_2 \cap (\sigma^+ \times \sigma^-)$.

We reduce the flow ℓ by ignoring the cycles and paths passing through node 0. In addition, the flow a carried by the path from source to sink $s \to i_1 \to i_2 \to \cdots \to i_r \to t$ is redirected to edge (i_1, i_r). As a result, we obtain a flow of value $a|\sigma^-|$ on the graph of the relation $\tilde{\Delta}_2$ restricted to $\sigma^+ \times \sigma^-$. This entails the existence of a matching of cardinality $|\sigma^-|$ in this graph, obtained by setting an upper capacity constraint of value 1 on each edge leaving the source s and entering the sink t (as a cut of capacity C on the network with capacity constraints $c_{i,j} \in \{1, \infty\}$ is a cut of capacity $a \times C$ on the same network with capacity constraints $a \times c_{i,j}$).

- (3) $\Rightarrow$ (4) is simply a rewording.
- (4) $\Rightarrow$ (1): Let $\phi : \sigma^- \to \sigma^+$, injective, such that $\forall k \in \sigma^-, (\phi(k), k) \in \tilde{\Delta}_2$. Given any ordering O of the negative argument set σ^-, we can build a sequence of alternatives of decreasing preference $e_0 := x, e_1, \ldots, e_{|\sigma^-|} \in V$ such that the kth statement (e_{k-1}, e_k) matches the criteria swap $(\phi(O_k), O_k) \in \tilde{\Delta}_2$:

$$N^{\neq}_{(e_{k-1}, e_k)} := \{\phi(O_k), O_k\} \, ; \; N^{=}_{(e_k, y)} := N^{=}_{(e_{k-1}, y)} \cup \{\phi(O_k), O_k\}$$

Thus, the sequence of sets $(e_k, y)^-$ decreases from σ^- to $\emptyset$, one element at a time, and the sequence of sets $(e_k \succsim y)^+$ also decreases from σ^+ to $\sigma^+ \backslash \phi[\sigma^-]$, one element at a time. If the set $\sigma^+ \backslash \phi[\sigma^-]$ is empty, $e_{|\sigma^-|} = y$, and the sequence $x = e_0, \ldots, e_{|\sigma^-|} = y$ is an explanation of $(x, y) \in \mathcal{N}_{\mathcal{P}}$ by preference swaps of order 2, of length $|\sigma^-|$. Else, $e_{|\sigma^-|} \neq y$ but $(e_{|\sigma^-|}, y)$ is a dominance statement, as its negative argument set is empty. Thus, the sequence $x = e_0, e_1, \ldots, e_{|\sigma^-|}, y$ is an explanation of $(x, y) \in \mathcal{N}_{\mathcal{P}}$ by preference swaps of order 2 and a dominance relation, of length $|\sigma^-| + 1$.

References

Aingworth, D., Chekuri, C., & Motwani, R. 1996. Fast estimation of diameter and shortest paths (without matrix multiplication). In *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '96, pp. 547–553

Bana e Costa, C. A., & Vansnick, J.C. (1995). General overview of the MACBETH approach. In Pardalos P. M., Siskos Y., & Zopounidis C., (Eds.) *Advances in Multicriteria Analysis*, pages 93–100. Dordrecht: Kluwer Academic Publishers

Bana e Costa, C. A., Lourencco J. C., Chagas, M. P. & Bana e Costa, J. C. (2008). Development of reusable bid evaluation models for the portuguese electric transmission company. *Decision Analysis*, 5(1):22–42

Boutilier, C., Brafman, R. I., Domshlak, C., Hoos, H. H., & Poole, D. (2004). Cp-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *J. Artif. Intell. Res. (JAIR)*, *21*, 135–191.

Brafman, R. I., Domshlak, C., & Shimony, S. E. (2006). On graphical modeling of preference and importance. *J. Artif. Intell. Res. (JAIR)*, *25*, 389–424.

Carenini, G., & Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence Journal*, *170*, 925–952.

Ch. Labreuche, Maudet N., & Ouerdane W. (2011). Minimal and complete explanations for critical multi-attribute decisions. In *Algorithmic Decision Theory (ADT)*, pp. 121–134, Piscataway, NJ, USA

Ch. Labreuche, Maudet, N., & Ouerdane, W. (2012). Justifying dominating options when preferences are incomplete. In *Proceedings of the European Conference on Artificial Intelligence*, *242*, pp 486–491, Montpellier, France, IOS Press

Eiter, T., & Gottlob, G. (1995). The complexity of logic-based abduction. *J. ACM*, *42*(1), 3–42.

Even, S., & Tarjan, R. E. (1975). Network flow and testing graph connectivity. *SIAM J. Comput.*, *4*(4), 507–518.

Fishburn, P.C. (1997). Cancellation conditions for multiattribute preferences on finite sets. In Mark, H. Karwan, J. S., & Jyrki W. (Eds.) *Essays In Decision Making*, pp. 157–167. Springer Berlin Heidelberg

Friedrich, G., & Zanker, M. (2011). A taxonomy for generating explanations in recommender systems. *AI Magazine*, *32*(3), 90–98.

Greco, S., Słowinski, R., Figueira, J., & Mousseau, V. (2010). Robust ordinal regression. In *Trends in Multiple Criteria Decision Analysis*, pp 241–284. Springer Verlag

Greco, S., Mousseau, V., & Słowinski, R. (2008). Ordinal regression revisited: Multiple criteria ranking with a set of additive value functions. *European Journal of Operational Research*, *191*, 416–436.

Hammond, J., Keeney, R., & Raiffa, H. (1998). Even Swaps: a rational method for making trade-offs. *Harvard Business Review*, 137–149

Herlocker, J. L., Konstan, J. A., & Riedl, J.(2000). Explaining collaborative filtering recommendations. In *Proceedings of the ACM conference on Computer Supported Cooperative Work*, pp. 241–250

Jacquet-Lagrèze, E., & Siskos, Y. (1982). Assessing a set of additive utility functions for multicriteria decision making: the UTA method. *European Journal of Operational Research*, *10*, 151–164.

Kazman, R., Klein, M., & Clements, P. (2000). *ATAM: Method for Architecture Evaluation*. TECHNICAL REPORT, CMU/SEI-2000-TR-004, http://www.sei.cmu.edu/reports/00tr004.pdf

Klein, D. A. (1994). *Decision analytic intelligent systems: automated explanation and knowledge acquisition*. Lawrence Erlbaum Associates.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*, volume 1: Additive and Polynomial Representations. Academic Press

Labreuche, Ch. (2011). A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence Journal*, *175*, 1410–1448.

Michell, Joel. (1988). Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology*, *32*(4), 466–473.

Nic, W. (2011). Computational techniques for a simple theory of conditional preferences. *Artificial Intelligence*, 175(7–8):1053–1091. Representing, Processing, and Learning Preferences: Theoretical and Practical Challenges.

Nunes, I., Miles, S., Luck, M., Barbosa, S., & Lucena, C. (2014) Pattern-based explanation for automated decisions. In *Proceedings of the 21st European Conference on Artificial intelligence*, pp. 669–674. IOS Press

O'Sullivan, B., Papadopoulos, A., Faltings, B., & Pu, P. (2007). Representative explanations for over-constrained problems. In *Proceedings of the 22nd national conference on Artificial intelligence*, pp. 323–328. AAAI Press

Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542 – 556. Special Issue On Intelligent User Interfaces.

Spliet, R., & Tervonen, T. (2014). Preference inference with general additive value models and holistic pair-wise statements. *European Journal of Operational Research*, *232*(3), 607–612.

Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2009). MoviExplain: a recommender system with explanations. In *Proceedings of the third ACM conference on Recommender systems (RecSys'09)*, pp. 317–320, New York, NY, USA, 2009. ACM.

Ulrich, J. (2004) Quickxplain: Preferred explanations and relaxations for over-constrained problems. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 167–172, Menlo Park, California, 2004. AAAI Press /The MIT Press.

C.2 Selection of articles related to Chapter 4

- Jean-Philippe Poli, Wassila Ouerdane, Regis Pierrard. Generation of Textual Explanations in XAI: the Case of Semantic Annotation. 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jul 2021, Luxembourg, Luxembourg. pp.9494

- Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. Comparing options with argument schemes powered by cancellation. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19), Macao,China. pp 1537-1543, 2019.

- Khaled Belahcène, Yann Chevaleyre, Nicolas Maudet, Christophe Labreuche, Vincent Mousseau, and Wassila Ouerdane. Accountable Approval Sorting. Proceedings of 27th International Joint Conference on Artificial Intelligence and 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018). Stockholm, Sweden. pp 70-76, 2018.

- Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau and Wassila Ouerdane. A Model for Accountable Ordinal Sorting. In proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-2017), Melbourne, Australia. pp 814-820, 2017.

- Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane. Justifying Dominating Options when Preferential Information is Incomplete. Proceedings of the 20th European Conference on Artificial Intelligence (ECAI'12), Montpellier, France. IOS Press, 242, pp.486-491, Frontiers in Artificial Intelligence and Applications. 2012.

- Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane. Minimal and Complete Explanations for Critical Multi-attribute Decisions. In Proceedings of the 2nd International Conference on Algorithmic Decision Theory (ADT'2011), Piscataway New Jersey, United States. Springer, Lecture Notes in Computer Science. pp.121-134, 2011.

Generation of Textual Explanations in XAI: the Case of Semantic Annotation

Jean-Philippe Poli*, Wassila Ouerdane† and Régis Pierrard*†
*Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{jean-philippe.poli, regis.pierrard}@cea.fr
†Université Paris-Saclay, CentraleSupélec, MICS, 91190, Gif-sur-Yvette, France
wassila.ouerdane@centralesupelec.fr

Abstract—Semantic image annotation is a field of paramount importance in which deep learning excels. However, some application domains, like security or medicine, may need an explanation of this annotation. Explainable Artificial Intelligence is an answer to this need. In this work, an explanation is a sentence in natural language that is dedicated to human users to provide them clues about the process that leads to the decision: the labels assignment to image parts. We focus on semantic image annotation with fuzzy logic that has proven to be a useful framework that captures both image segmentation imprecision and the vagueness of human spatial knowledge and vocabulary. In this paper, we present an algorithm for textual explanation generation of the semantic annotation of image regions.

Index Terms—Explanation, natural language generation, semantic annotation, fuzzy constraint satisfaction problems

I. INTRODUCTION

Semantic image annotation is the ability for a computer to label images or image regions. It is a task of paramount importance with the daily production of images in all the domains (e.g. medicine, surveillance).

In this field, deep learning has enabled to build models that can efficiently classify images and recognize objects. Sometimes, these models can even top human capabilities on several specific tasks [1]. For some critical applications of Artificial Intelligence (AI), performance is not the only criterion to optimize [2]. Such applications may require a relative understanding of the logic performed by the AI. In other words, the end-user would like to get a response to the question "Why ?" [3]

For semantic annotation, Constraint Satisfaction Problems (CSP) have been successfully applied to geometrical figure annotation [4] and region labelling from a model [5]. Vanegas et al. extended these previous works to fuzzy constraint satisfaction problems (FCSP) to involve fuzzy spatial relations and illustrate their approach with an automatic interpretation of Earth observation images [6]. Since CSP and FCSP are interpretable models and the process of solving is also interpretable and explainable, this kind of approaches are good candidates for explainable semantic annotation of images. Pierrard et al. [7] propose algorithms to extract automatically relevant fuzzy spatial relations for image annotation from

a few learning images whose regions are segmented and labelled. The appropriate relations are then used to constitute a FCSP for annotating areas of an image or a rule base to classify the image.

In this paper, we focus on the generation of a textual explanation of the semantic annotation in the context of [7]. Given a solution of such FCSPs and the degree of satisfaction of all the involved constraints, we propose and evaluate two algorithms to extract clues of the reasoning and to order the pieces of the explanation efficiently.

The paper is structured as follows. In section II, fuzzy spatial relations, constraint satisfaction problems and their solving are described. Next, sections III and IV are devoted to describe the methods for generating explanation of semantic annotation. Then, the two approaches are evaluated and compared in section V. Finally, we draw some conclusions and perspectives in section VI .

II. BACKGROUND

A. Fuzzy Spatial Relations

The fuzzy logic framework allows using words instead of numbers during computations and also during problem formalization. Indeed, relations are represented by a linguistic description that can be directly used in the explanation [7].

Many fuzzy spatial relations have been studied in the literature [8]. For instance, Vanegas considers three types of spatial relations: topological, metric and structural relations [6]. The two first types are often used in computer vision. We can cite for instance the RCC8 framework that defines relations between regions and their fuzzy counterparts that have been introduced in [9], [10]. Bloch introduced a framework based on fuzzy morpho-mathematics to evaluate fuzzy spatial relations [8]. In particular, metric directional relations can be expressed based on the fuzzy dilation operator.

Without loss of generality, in the remainder of this paper, we use specifically directional, distance and symmetry relations. Directional and distance relations [8] are computed as a fuzzy landscape and assessed using a fuzzy pattern matching approach [11]. The symmetry relation [12] we use consists in finding the line that maximizes a symmetry measure between two objects (regions). Since this measure is not differentiable, a direct search method is used to solve this optimization problem, such as the downhill simplex method.

B. Fuzzy Constraint Satisfaction Problems

A constraint satisfaction problem (CSP) consists in assigning some values to a set of variables that must respect a set of constraints.

An extension of CSP to the fuzzy logic framework to deal with imprecise parameters and flexible constraints is presented in [13]. This is called a fuzzy constraint satisfaction problem (FCSP). A FCSP is defined by:

- A set of variables $X = \{x_1, ..., x_n\}$,
- A set of domains $D = \{D_1, ..., D_n\}$ such as D_i is the range of values that can be assigned to x_i,
- A set of flexible constraints $C = \{c_1, ..., c_p\}$. Each constraint c_k is defined by a fuzzy relation R_k and by the set of variables V_k that are involved in it.

To solve a FCSP, the backtracking algorithm is applied. It starts with an empty set of instantiations and selects a variable $x \in X$ to instantiate. Then, it finds a value in the domain of X that maintains the consistency of the current instantiation, regarding the set of constraints C. The steps are repeated until all the variables are instantiated. When a variable x has no more value to test, the algorithm backtracks and tries the next value of the previously instantiated variable.

An instantiation that is consistent and complete is a solution. One solution of the FCSP is evaluated by its degree of consistency. Given a solution γ, its degree of consistency [6] is:

$$\text{cons}(\gamma) = \min_{c_k \in C} \mu_{R_k}(\gamma_{|V_k}) \tag{1}$$

where $\gamma_{|V_k}$ is the projection of γ on V_k and μ_{R_k} the membership function representing R_k.

This consistency degree also enables to compare different solutions so that the best one can be extracted.

To improve the performance of the backtracking algorithm, [6], [13] have adapted the AC-3 algorithm of crisp CSP that prunes the domains, discarding values that are inconsistent with the current instantiation.

C. Image Annotation with FCSP

When dealing with image annotation, the set of variables X corresponds to the objects we would like to instantiate. The variables share the same domain D that represents the regions in the image that we get after segmentation. Thus, $|X| \leq |D|$. The constraints in C are defined by fuzzy relations: some of them can deal with groups of objects [6].

This can solve specific annotation problems in which the objects to annotate and the labels are known (even if they are automatically detected, by a segmentation for instance). The intuition behind is that such annotation problem can be combinatorial and the labels are affected accordingly to each other, by opposition with individually like in classical approaches.

In [7], this approach was applied to organ annotation in medical images, with a focus on automatically generating the FCSP from few data. In the remainder of this paper, we will take this work as an illustration with an automatic generated FCSP.

To generate our explanations, the algorithms we propose in this work (Algo 1 and Algo 2) take as input a trace $T = \langle P, s, \bar{C} \rangle$ of the execution of the solving algorithms. T is composed of:

- $P = \langle X, D, C \rangle$ is a FCSP.
- s, a chosen solution among all the solutions of P, for instance the best one regarding the degree of consistency. s contains the assignment for each variable in X.
- $\bar{C}$, the set of degrees of satisfaction of each $c \in C$.

D. Surface Realization

In linguistics, a realization consists in generating a *surface form*, which is a correct sentence in a given natural language, from a more abstract representation, in which the different components such as the subject or the verb are specified. Therefore, a *surface realizer* is a system that is able to take an abstract semantic representation as an input to generate a syntactically-correct sentence.

In this work, we rely on SimpleNLG [14] for performing this task. This realization engine provides an API that is easy to use and complete enough for the kind of explanation we would like to generate. We do not explain here how we use it (e.g. the function calls). We will just describe the form of the sentences.

III. Complete Textual Explanation Generation

In this section, we present a first algorithm for explanation generation in natural language.

A. Algorithm

Algo1 uses all the constraints of the FCSP and turns them into sentences. The vocabulary of relations contains: *to the left of*, *to the right of*, *below*, *above*, *close to*, *symmetrical to* and *stretched*. That makes 6 binary and one unary relations. We note that $\bar{c}_x$ is the complement of x in the scope of c, and the *moderator* is selected among those cited in table I according to the satisfaction of c. This idea is inspired from [15].

Moderator	Degree of satisfaction
very high	from 0.9
high	from 0.7 to 0.9
average	from 0.4 to 0.7
moderate	from 0.2 to 0.4
low	from 0 to 0.2

TABLE I
SIMPLIFIED CONFIDENCE SCALE

B. Results

In this work, the FCSP has been extracted automatically from few images from the Visceral dataset[1]. Figure 1 shows one of the image and different organs of interest.

The segmentation has been obtained automatically and the regions were given an identifier in an arbitrary order. Thus, in this first approach, items are not sorted. However, for the sake

[1] http://www.visceral.eu/

Algorithm 1: Complete Explanations Generation

Input: a trace $T = \langle P, s, \bar{C} \rangle$
Output: a complete textual explanation
1 **foreach** *unprocessed variable* $x \in X$ **do**
2     $v \leftarrow$ value of x in s
3   Create a sentence of the form: "Region v is annotated as x with a *moderator* confidence because:"
4   **foreach** *constraint* $c \in C$ *involving* x *in its scope* **do**
5   **if** *x is the first variable in the scope of c* **then**
6   Generate a sentence of the form: "it is $c\ \bar{c}_x$" (eventually, for each variable $x' \in \bar{c}_x$, indicate the associated $v' \in s$)
7   **else**
8   Generate a sentence of the form: "$\bar{c}_x$ is/are $c\ x$"
9   **end**
10   **end**
11 **end**

of comprehension of this article, we numerated ourselves the organs, from left to right and top to bottom.

We consider the solution of such a FCSP for Figure 1 with the highest degree of consistency.
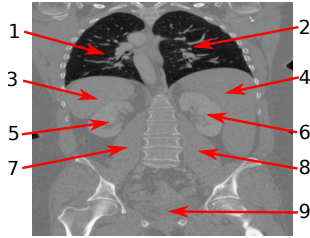
Fig. 1. Backward MRI image with different regions to annotate

The result, as it can be seen in figure 2 is obviously a long but complete explanation.

In the next, we investigate the possibility to shorten this explanation. Thus, the next section is dedicated to describe a second algorithm to generate a more concise explanation.

IV. CONCISE TEXTUAL EXPLANATION GENERATION

A. Cognitive Science Considerations

Cognitive science has largely studied the way Humans represent a scene or scan images. Thus, it seems natural to consider those insights to create an explanation.

Zwaan et *al.* present more than a decade of studies about situation model, i.e. a mental representation of affairs [16]. They highlight the difficulty to describe correctly a spatial scene with language, because of the difference between its dimensionality and the dimensionality of space. For instance, if one describes a room in a circular way, the first and the last objects are far from each other in the description but close in the room. This also shows the importance of the order in which the parts of the scene have to be described.

This leads us to the studies about image scanning [17], which is related to the mental representation of a scene or an image. Authors of [18] state that the visual images preserve the metric spatial information. This implies that starting from a

Region 1 is annotated as the left lung with a high confidence **because**:

- it is *completely to the left of* region 2 (annotated as the right lung by the model),
- region 2 (right lung) is *completely to the right of* region 1,
- it is *above* region 3 (spleen),
- region 3 (spleen) is *completely below* region 1,
- it is *above* region 7 (left psoas),
- region 7 (left psoas) is *completely below* region 1,
- region 5 (left kidney) is *completely below* region 1.

Region 2 is annotated as the right lung with a very high confidence **because**:

- it is *completely to the right of* region 1 (left lung),
- region 1 (left lung) is *completely to the left of* region 2,
- region 3 (spleen) is *to the left of* region 2,
- region 4 (liver) is *below* region 2,
- it is *above* region 8 (right psoas),
- region 8 (right psoas) is *completely below* region 2,
- region 6 (right kidney) is *completely below* region 2,
- region 9 (bladder) is *below* region 2.

Fig. 2. Extract of an explanation for an annotation with the complete approach. The complete one can be found in [7].

focus point, subjects need more and more time to mentally visualize the information when going further to this focus point. Other works study the difficulties of subjects to represent a scene if the description is too long and if the description is too precise [19], [20]. Another difficulty is the direction of reading: [21] indicates that it affects the description of a scene.

The studies about image scan paths bring also good information. The attention of subjects is classically attracted by focus points. In image understanding, this is called salient objects and [22] gives a comprehensive review on their automatic detection. Nevertheless, cognitive science warns of the difficulty of defining saliency because it can be context-dependent, or due to the singularity of an object, of the user's goal, etc. However, when a same subject watches the same picture, the scan paths may be different [23]: thus, the scan path does not depend only on the objects in the image. If several similar pictures are presented, the scan path can also be more and more efficient [23].

Finally, the Gestalt psychologists [24] studied the cognitive issues of visual perception, in particular the shape of objects. The 7 Gestalt principles concern figure-ground, similarity, proximity, common region, continuity, closure and focal point of images. They are particularly useful in design, but give some insight about how objects are perceived. In particular, they recommend to group objects that are similar or that share properties.

This short overview of cognitive science helped us to design our explanation strategy.

B. Principles

The previous subsection gives raw information from the cognitive science. The idea of our approach is to improve the previous version of the generation of explanation from a FCSP by considering cognitive science insights. We thus observe these principles:

- *Sorting*: the order of the results has an importance. It is important to start with regions in images that are salient, and then, regarding the recommendations of cognitive science papers, use diagonals and increasing distances to select the next results. The spiral order is not recommended.
- *Saliency*: the saliency is a difficult concept that can be context-dependent. *A minima*, one can select the biggest object or a group of objects as focus point.
- *Symmetry*: a pair of objects that are symmetrical must be grouped.
- *Priority*: we must select the most satisfied constraints first.
- *Associativity*: some relations are associative (e.g. "to the left of") and explainees can immediately infer it, so we must use that to reduce the number of constraints involved in the explanations.
- *Locality*: if possible, we will use first the constraints with the closest regions in the image.

Moreover, an explanation must somehow indicate how the task has been achieved. In our case, the solving of a FCSP is quite simple to explain since the algorithm searches for the values of the variables such as the constraints are satisfied. However, it makes the explanation more complicated when constraints are not all unary, since these assignments are dependent from each other. Indeed, for instance, a binary constraint will force the assignment of two variables together. In the case of semantic annotation or classification, the constraints are relations so that it is a little bit simpler than, for instance, quadratic constraints.

Another point is that we are selecting a maximum number of constraints for each variable, such that there is no correlation between these constraints: for instance, the values of "to the left of" and "to the very left of" may be correlated and so we do not want to use them at the same time for the same variable because they are redundant. We use mutual information to detect this correlation.

In the next subsection, we introduce an algorithm that considers those different principles.

C. Algorithm

Algo 2 presents the algorithm to generate concise explanation for semantic annotation.

The explanation starts with a general sentence that indicates the global confidence about the annotation based on the degree of consistency of the solution (line 1). The algorithm then selects the region from the segmentation that is the most salient (line 2). Regarding this object, the image is divided into four quadrants. The explanation will start with the most salient region, then with the other objects in the same quadrant, then quadrant by quadrant, in the clockwise order. This order is materialized in an ordered set X' (lines 3-4).

For each variable in X', the algorithm has to select at most N_{max} constraints to justify the explanation. The constraints are chosen regarding not only their level of satisfaction (that must be the highest as possible not to overload the text with moderators), but also their mutual link and the proximity with the other variables (lines 5-12).

The mutual link between relations is a tricky part. We use a knowledge graph about the relations as proposed in [7]. Such a graph emphasizes different links between two relations r_1 and r_2, like $r_1 \implies r_2$, $\neg r_1 \implies r_2$, but also symmetry. Symmetry is important not to use twice the same constraint. Let o_1 and o_2 be two objects in the image, and r a symmetrical relation, if $o_1 \ r \ o_2$ is used in a sentence, we cannot use $o_2 \ r \ o_1$ anymore.

Then, the algorithm looks for grouping constraints such as "is symmetrical to" that constitutes a pair of variables (line 9). Indeed, the previous section highlights that groups of objects must be treated together. Thus, the other variables in the scope of this constraint must be processed just after (line 10).

Algorithm 2: Concise Explanation Generation

Input: a trace $T = \langle P, s, \bar{C} \rangle$
Output: a concise textual explanation

1 Write a sentence to introduce the result and the global confidence
2 Select f the variable in s region that is the focus point in the image
3 From the center of f, divide the image into 4 quadrants $Q_1, \ldots, Q_4$
4 $X' =$ set of variables $x \in s$ sorted by quadrant
5 **while** $X' \neq \emptyset$ **do**
6 $\quad$ $x \leftarrow pop(X')$
7 $\quad$ $S \leftarrow$ Select N_{max} constraints c_i that are not linked in the knowledge graph and with maximal degrees of satisfaction
8 $\quad$ Write the sentence "x is c_1, ..., and is $c_{j \leq N_{max}}$"
9 $\quad$ **if** x *involves a grouping constraint* c **then**
10 $\quad\quad$ Move all variables in scope of c to the beginning of X'
11 $\quad$ **end**
12 **end**

D. Results

In this work, we define the focus point as the biggest object (in terms of area). We set $N_{max} = 2$.

For the same example (see Figure 1), and the same solution s, the result is shown in Figure 3.

Most of the constraints are linked in the knowledge graph, because we used mainly directional relations like "to the right of" and "to the left of". This explains why we rarely reach N_{max} constraints.

The result is obviously shorter, and seems easier to read. The quadrant imposes an order for the description of each organ. The explanation seems less redundant thanks to the selection of the constraints.

"This is the annotation of the given image (with a very high confidence). The right lung (region 2) is *symmetrical to* the left lung (region 1) and *above* the liver (region 4).
The liver (region 4) is *at the right of* the right kidney (region 6) and *at the right of* the right psoas (region 8).
The right psoas (region 8) is *above of* the bladder (region 9) and is *symmetrical to* the left psoas (region 7).
The left psoas (region 7) is *below* the left kidney (region 5).
The spleen (region 3) is *above* the left kidney (region 5) and is *below* the left lung (region 1)."

Fig. 3.   Concise explanation produced by Algorithm 2

The next section is dedicated to the evaluation of both types of explanation.

V. Evaluation and Discussion

To compare the two approaches, we evaluated both of them. In this aim, we use the questionnaire presented in [25]: it is based on 17 questions organized in 3 categories: natural language, human-computer interaction and content and form. Each question is evaluated with a Likert scale (from 1 "strongly disagree" to 5 "strongly agree"). Our panel consists in 40 respondents, with 20 medical staff members (medical doctors, surgeons, nurses, radiologists), the other half being computer scientists (6) and other various non-medical professionals (14). To decrease the medical staff's amount of time dedicated to the questionnaire, we selected only 14 questions out of the 17 initial ones that will allow comparing the both approaches. We removed the questions about the grammar and the one that indicates if the explanation made a respondent change his mind. Because of the lack of space, figure 4 highlights the answers to few questions.

Both explanations are comparable in terms of syntax correctness (87% for approach 1 and 95% for approach 2), of reasoning comprehension (67.5% agree for approach 1, 60% for approach 2), and of uncertainty communication (62.2% for approach 1, 65% for approach 2). "Reasoning comprehension" indicates if the respondents can infer about the reasoning process when they read the explanation. The "uncertainty communication" criterion evaluates the ability of the explanation to tell the user at which point the decision can be trusted. In our case, it is achieved by the translation of the constraints satisfaction into sentence parts like "with a very high confidence". These facts show that not all the people understood how the algorithm annotates the organs and understood why the algorithm was not confident in all the cases.

For all other comparisons, the second approach outperforms the first approach. 19 persons found that the first explanation was too long whereas only 1 respondent was concerned by the length of the second explanation. Respondents found the first explanation repetitive (87.5%) and hard to read (72.5%), whereas only respectively 22.5% and 10% of the panel agree with these facts for the second one. Only 32.5% of the respondents found the order of the items in the explication suitable for explanation 1 versus 72.5% for the second explanation.

Both explanations make the respondents think they can trust the automatic labelling (55% for first approach and 65% for the second one).

These results confirm the advantages of the second algorithm.

First, it is important to note that these algorithms are not domain-specific. Indeed, the relations are generic in the sense that they could be used in another domain (such as satellite image annotation). They also manipulate image regions, and have no clue they represent organs. However, the labels that are used are organ names, because we want a semantic annotation. We do not use external domain knowledge, for instance to replace the word "region" by "organ" on the explanation, or to use a more technical vocabulary.

The results show that the order of the items inside an explanation are important for the end users. Conciseness seems to be a criterion of paramount importance too.

The questionnaire invited also the respondents to write comments after each type of explanation. Most of the medical staff felt uncomfortable with the fact that the MRI image was taken from the back. Nevertheless, no one declared the explanation was wrong: maybe it can have an impact on the confidence of the users in the AI.

One of the medical respondent said it could be useful to use the spine as main region and use it for the labelling of the other regions. This idea emphasizes the importance of saliency: indeed, in such an image, we can see the spine first because it is whiter and central. Unfortunately, in the segmentation we use, bones are not considered.

Finally, we also made a comparison between the medical respondents and the others, but the results do not differ significantly.

VI. Conclusion and Perspectives

In this paper, we presented our work on the generation of textual explanations of image annotation. The first part provides a form of explanation that was not pertinent for humans. The second part is an improvement of the first one that generates a more concise explanation. It relies on a more sophisticated selection of the constraints that are used in the explanation, based on cognitive science principles.

This work also shows the importance of realizers for explainable AI: although it is not the goal of this work, using synonyms or different sentence structures to break the monotony of the explanations can help. However, the survey we presented shows that most participants are convinced by the explanations and they understand the logic of the model.

What we observe is that to develop a model, then an algorithm to extract relevant clues and finally improve realizers involve too many fields and is difficult to manage. In our future work, we are thinking of the separation of these tasks.

Strongly disagree ▪ Disagree ▪ Neutral ▪ Agree ▪ Strongly agree

-40 -30 -20 -10 0 10 20 30 40

Explanation #1's language is correct
Explanation #2's language is correct
Explanation #1 is easy to read
Explanation #2 is easy to read
Explanation #1 is convincing
Explanation #2 is convincing
Explanation #1 allows understanding how the result was inferred
Explanation #2 allows understanding how the result was inferred
Explanation #1's length is adequate
Explanation #2's length is adequate
Explanation #1 is not repetitive
Explanation #2 is not repetitive
It was easy to reach the end of the explanation #1
It was easy to reach the end of the explanation #2
Explanation #1's items order is natural
Explanation #2's items order is natural
Explanation #1 helps trusting the results
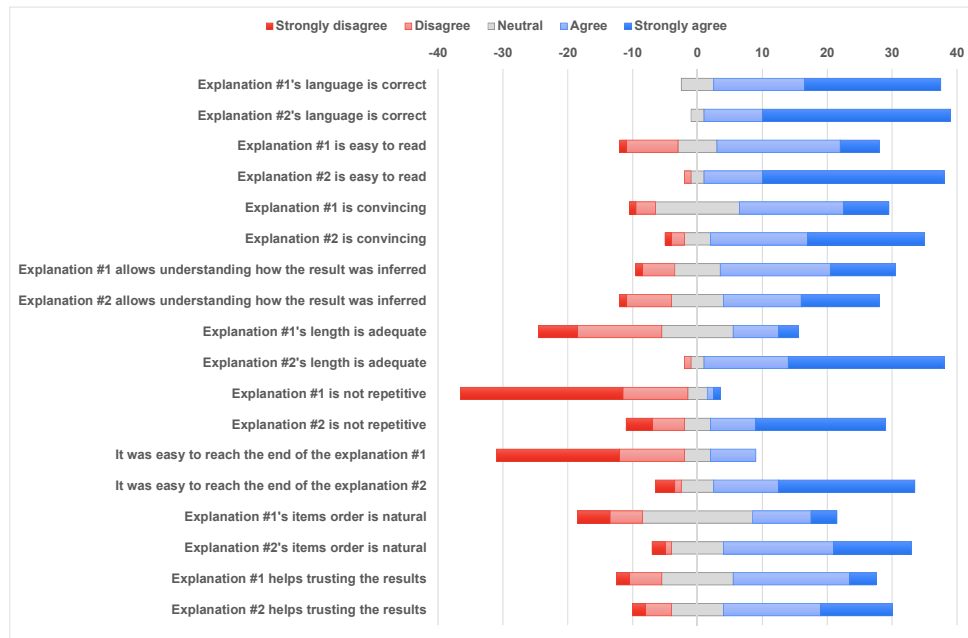Explanation #2 helps trusting the results

Fig. 4. Highlights from the survey

REFERENCES

[1] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. D. Fauw, and S. Shetty, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 1 2020.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," in *eprint arXiv:1702.08608*, 2017.

[3] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 2679, pp. 1–38, 2019.

[4] D. Waltz, "Understanding line drawings of scenes with shadows," in *The Psychology of Computer Vision*. McGraw-Hill, 1975.

[5] A. Deruyver, Y. Hodé, and L. Brun, "Image interpretation with a conceptual graph: Labeling over-segmented images and detection of unexpected objects," *Artificial Intelligence*, vol. 173, no. 14, pp. 1245 – 1265, 2009.

[6] M. C. Vanegas Orozco, "Spatial relations and spatial reasoning for the interpretation of Earth observation images using a structural model." Theses, Télécom ParisTech, Jan. 2011.

[7] R. Pierrard, J.-P. Poli, and C. Hudelot, "Spatial relation learning for explainable image classification and annotation in critical applications," *Artificial Intelligence*, vol. 292, p. 103434, 2021.

[8] I. Bloch, *Fuzzy Models of Spatial Relations, Application to Spatial Reasoning*. Springer Berlin Heidelberg, 2013, pp. 51–58.

[9] Yongming Li and Sanjiang Li, "A fuzzy sets theoretic approach to approximate spatial reasoning," *IEEE Transactions on Fuzzy Systems*, vol. 12, no. 6, pp. 745–754, Dec 2004.

[10] S. Schockaert, M. D. Cock, C. Cornelis, and E. E. Kerre, "Fuzzy region connection calculus: Representing vague topological information," *International Journal of Approximate Reasoning*, vol. 48, no. 1, pp. 314 – 331, 2008, special Section: Perception Based Data Mining and Decision Support Systems.

[11] M. Cayrol, H. Farreny, and H. Prade, "Fuzzy pattern matching," *Kybernetes*, vol. 11, no. 2, pp. 103–116, 1982.

[12] O. Colliot, "Représentation, évaluation et utilisation de relations spatiales pour l'interprétation d'images. application à la reconnaissance de structures anatomiques en imagerie médicale," Ph.D. dissertation, Télécom ParisTech, 2003.

[13] D. Dubois, H. Fargier, and H. Prade, "Possibility theory in constraint satisfaction problems: Handling priority, preference and uncertainty," *Applied Intelligence*, vol. 6, no. 4, pp. 287–309, 1996.

[14] A. Gatt and E. Reiter, "Simplenlg: A realisation engine for practical applications," in *Proceedings of the 12th European Workshop on Natural Language Generation*, 2009, pp. 90–93.

[15] D. V. Budescu, H.-H. Por, and S. B. Broomell, "Effective communication of uncertainty in the IPCC reports," *Climatic change*, vol. 113, no. 2, pp. 181–200, 2012.

[16] R. A. Zwaan and G. A. Radvansky, "Situation models in language comprehension and memory." *Psychological Bulletin*, vol. 123, no. 2, p. 162, 1998. [Online]. Available: https://app.dimensions.ai/details/publication/pub.1041143270

[17] G. Borst, S. Kosslyn, and M. Denis, "Different cognitive processes in two image-scanning paradigms," *Memory & cognition*, vol. 34, pp. 475–90, 05 2006.

[18] S. Kosslyn, T. Ball, and B. Reiser, "Visual images preserve metric spatial information: Evidence from studies of image scanning," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 4, no. 1, pp. 47–60, 2 1978.

[19] M. J. Farah and S. M. Kosslyn, "Structure and strategy in image generation," *Cognitive Science*, vol. 5, no. 4, pp. 371 – 383, 1981.

[20] M. Denis, M.-R. Goncalves, and D. Memmi, "Mental scanning of visual images generated from verbal descriptions: Towards a model of image accuracy," *Neuropsychologia*, vol. 33, no. 11, pp. 1511 – 1530, 1995, the Neuropsychology of Mental Imagery.

[21] A. Román, A. Fathi, and J. Santiago, "Spatial biases in understanding descriptions of static scenes: The role of reading and writing direction," *Memory & cognition*, vol. 41, 01 2013.

[22] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, pp. 117–150, 2014.

[23] D. Noton and L. Stark, "Scanpaths in eye movements during pattern perception," *Science*, vol. 171, no. 3968, pp. 308–311, 1971.

[24] R. Luccio, *Gestalt problems in cognitive psychology: Field theory, invariance and auto-organisation*. Springer Berlin Heidelberg, 1993, pp. 1–19.

[25] I. Baaj and J. Poli, "Natural language generation of explanations of fuzzy inference decisions," in *2019 IEEE International Conference on Fuzzy Systems*. IEEE, 2019, pp. 1–6.

Comparing Options with Argument Schemes Powered by Cancellation

Khaled Belahcene[1] , **Christophe Labreuche**[2] , **Nicolas Maudet**[3*] , **Vincent Mousseau**[4] and **Wassila Ouerdane**[4]

[1] Nutriomics, Sorbonne Université, INSERM, France

[2] Thales Research and Technology, Palaiseau, France

[3] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

[4] MICS, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

khaled.belahcene@polytechnique.org, christophe.labreuche@thalesgroup.com, nicolas.maudet@lip6.fr, {vincent.mousseau, wassila.ouerdane}@centralesupelec.fr

Abstract

We introduce a way of reasoning about preferences represented as pairwise comparative statements, based on a very simple yet appealing principle: cancelling out common values across statements. We formalize and streamline this procedure with argument schemes. As a result, any conclusion drawn by means of this approach comes along with a justification. It turns out that the statements which can be inferred through this process form a proper preference relation. More precisely, it corresponds to a necessary preference relation under the assumption of additive utilities. We show the inference task can be performed in polynomial time in this setting, but that finding a minimal length explanation is NP-complete.

1 Introduction

In his famous letter to his friend Joseph Priestley, Benjamin Franklin suggested a procedure to decide upon difficult decision cases: draw two columns, list pros and cons, and delete (sets of) arguments from both sides when they are of "equal weight". It is remarkable that Franklin's "Moral Algebra" is sometimes seen as a pioneer technique to both argumentative approaches [Toulmin, 1958] which aims at formalizing, visualizing (and eventually criticizing) reasoning steps; as well as techniques to elicitate and reason about preferences based on trade-offs (*even swaps*, [Hammond *et al.*, 1998]). The bipolar nature of his algebra also proved to be influential in KR in general [Dubois *et al.*, 2008; Bouyssou *et al.*, 2009]. In this paper we build on the legacy of this approach, by relying on its core principle of cancellation, but without considering the weighting of different attributes – that is, only similar values can be crossed. We consider comparative preference statements whereby a user expresses unambiguously holistic judgments over alternatives described according to several points of view. From a set of such comparative statements, we wish to maintain the set of all valid consequences in order to make new inferences (under the form of further holistic comparative statements), and at the same time keep track of the reasoning steps involved. Thus, our objective is to know, for any preference query, *whether* it can be derived, but also *how* it can be derived.

We begin by introducing informally, through an example, a way of reasoning about preferences under the form of pairwise preference statements, then we propose a research agenda concerning this reasoning engine, and we outline the remainder the paper.

Example 1. *Hotels are compared according to the points of view of comfort, offer of a restaurant, commute time and cost. We are given* monotonicity *conditions according to each point of view (shared by all stakeholders), e.g. the larger the comfort the better; it is better to have a restaurant; the smaller the commute time the better; and the smaller the price the better. We are also given the following* preference information*:*

- π_1 : *a hotel with features (4*, no, 15 min, 180 \$) is preferred to a hotel with features (2*, yes, 45 min, 50 \$).*

- π_2 : *a hotel with features (4*, no, 45 min, 50 \$) is preferred to a hotel with features (4*, yes, 15 min, 100 \$).*

Monotonicity along each point of view allows for inferring comparative statements outside of the knowledge base.

Example 2. *(Ex. 1 continued) From π_1, thanks to the monotonicity w.r.t to cost, we can deduce that (4*, no, 15 min, 180 \$) is preferred to (2*, yes, 45 min, 80 \$).*

Reasoning *ceteris paribus* offers another venue to extend our knowledge about valid preferences.

Example 3. *(Ex. 2 continued) In π_2, both hotels share the same comfort rating 4*, and we propose to interpret this statement as: comfort being equal, we prefer (no, 45 min, 50 \$) to (yes, 15 min, 100 \$). Taking value 2*, we obtain for instance that (2*, no, 45 min, 50 \$) is preferred to (2*, yes, 15 min, 100 \$).*

These two principles are too weak to deduce many entailments from the preference information, and therefore will not allow comparing many alternatives. We therefore propose a way to combine several preference information statements. The notion of ceteris paribus reasoning can be generalized

---

*Contact Author

throughout statements by cancelling a similar value appearing in the left hand side (LHS) of a statement and in the right hand side (RHS) of another statement.

Example 4. *(Ex. 3 continued) For instance, the cost value 50 \$ appears both in the LHS of π_2 and the RHS of π_1. This extended principle can be used to infer new statements, as illustrated in the following table. The first three lines of the table introduce the* premises*: the preference information statements π_1 and π_2, as well as the ceteris paribus monontonicity statement $d_{¶}$ according to which, everything else being equal, a hotel with a restaurant is at least as good as a hotel without one. In each column representing a feature, we strike out individual values appearing simultaneously on the LHS and the RHS—when a value is repeated, we are careful to strike out as many values from the LHS as from the RHS. At the end, we notice that there is only one value left in each column, that we report on the last line of the table—the* conclusion*—forming what we consider a valid preference statement inferred from the premises.*

As: 4*	no	~~15 min~~	180 \$	π_1	2*	~~yes~~	45 min	~~50 \$~~
	~~no~~	45 min	~~50 \$~~	π_2		yes	~~15 min~~	100 \$
	~~yes~~			$d_{¶}$		~~no~~		
So, 4*	no	45 min	180 \$		2*	yes	45 min	100 \$

We are now in a position to set out several research questions concerning the procedure we just informally described, that we shall address in this paper:

- *Modeling (see Section 2).* To what extent this procedure can be formalized into a reasoning model?

- *Templating (see Section 2).* Can the template for presenting the arguments supporting a claim be streamlined? How can these bundles be efficiently validated?

- *Properties (see Section 3).* The set of statements that can be inferred from a given preference information form a binary relation between alternatives. What are the properties of this relation? Importantly, is it a proper *preference relation*?

- *Inference (see Section 3).* Is there an efficient way to assess if a given pairwise preference statement can be inferred from a given preference information?

- *Explanations (see Section 4).* Given a pairwise preference statement, is it possible to find a cognitively simple certificate supporting or informing its validity?

- *Critique (see Section 5).* This reasoning engine is built upon several fundamental assumptions, that need to be discussed.

2 The Reasoning Model

We address in this section the first research question, namely: can the intuition presented in the introduction be formalized?

2.1 Features and Alternatives

We consider a set N of points of view, each one $i \in N$ expressed by a feature taken in the set $\mathbb{X}_i$. Alternatives are described as tuples of features, and belong to the Cartesian product $\mathbb{X} = \prod_{i \in N} \mathbb{X}_i$.

For an alternative $x \in \mathbb{X}$ and a point of view $i \in N$, we denote by x_i the evaluation of x according to i. For any nontrivial subset of points of view $A \subset N$ and any two alternatives $a, b \in \mathbb{X}$, we denote $a_{-A}b_A$ the (fictitious) alternative which is equivalent to a according to each point of view not in A, and equivalent to b according to the points of view in A.

2.2 Preference Information

We are interested in providing a principled way of reasoning that allows us to infer preference and answer *preference queries* of the type 'is alternative a preferred to alternative b?'. The reasoning shall be based on *preference information*, coming in two distinct flavors:

- *explicit pairwise statements* $\mathcal{P} \subset \mathbb{X}^2$, where $(x, y) \in \mathcal{P}$ means that x is at least as good as y for the decision maker;

- *implicit dominance*—we assume that each feature set corresponding to a point of view $i \in N$ is totally ordered by a relation $\succsim_i \subset \mathbb{X}_i^2$, and we denote $\mathcal{D} := \prod_{i \in N} \succsim_i$, the Pareto *dominance* relation between alternatives stemming from the ordering of each feature set, i.e. $\forall x, y \in \mathbb{X}, x\mathcal{D}y \iff \forall i \in N, x_i \succsim_i y_i$.

2.3 Cancellation Axioms

The inductive principle based on ceteris paribus sketched in Ex. 3 can be formalized thanks to the concept of cancellation. The cancellative axioms are well-known in the preference literature [Krantz *et al.*, 1971; Wakker, 1989], and we briefly recall their definition.

Definition 1 (First-order cancellation)**.** *For all $A \subset N$, with $A \neq \varnothing$ and all $x, y, z, z' \in \mathbb{X}, x_{-A}z_A \succsim y_{-A}z_A \Rightarrow x_{-A}z'_A \succsim y_{-A}z'_A$.*

In Ex. 3, according to the first-order cancellation, π_2 implies that $(2^*, \text{no}, 45 \text{ min}, 50 \text{ \$})$ is preferred to $(2^*, \text{yes}, 15 \text{ min}, 100 \text{ \$})$.

We also have seen in the introduction cancellation *across* preference statements. It can be formalized in the following definition.

Definition 2 (High-order cancellation)**.** *Consider $m + 1$ alternatives $x^{(0)}, \ldots, x^{(m)}$ in $\mathbb{X}$. Let $y^{(0)}, \ldots, y^{(m)}$ be $m + 1$ alternatives in $\mathbb{X}$ such that, for every point of view $i \in N$, $(y_i^{(0)}, \ldots, y_i^{(m)})$ is a permutation of $(x_i^{(0)}, \ldots, x_i^{(m)})$. Then, $[x^{(k)} \succsim y^{(k)}, \forall k \in \{1, \ldots, m\}] \Rightarrow y^{(0)} \succsim x^{(0)}$.*

In order to conveniently represent concatenations of premises, maybe with repetition, modulo permutation, we represent the tuples of alternatives or values as *multisets*. The multiset containing the elements z_1, repeated m_1 times, $\ldots, z_k$ repeated m_k times has support $\{z_1, \ldots, z_k\}$, cardinality $\sum m_j$ and is denoted $\langle z_1 : m_1, ..., z_k : m_k \rangle$.

2.4 The Syntactic Cancellative Argument Scheme

We formalize the way of reasoning about preference statements illustrated in the introduction through an *argument scheme* [Walton, 1996], an operator tying *premises* satisfying some conditions, to a *conclusion*. This scheme is closely

related to the *high-order cancellation* axiom described previously. We slightly alter it in order to allow for a repetition of the conclusion (we defer an example and the discussion of the importance of this alteration to Section 3.5).

Definition 3 (Syntactic cancellative argument scheme). *Given two positive integers $m \geq n$, and a pair of alternatives $(x, y) \in \mathbb{X} \times \mathbb{X}$, we say the multiset of pairs of alternatives $\langle (a^{(1)}, b^{(1)}) : r_1, \ldots, (a^{(k)}, b^{(k)}) : r_k \rangle \in (\mathbb{X} \times \mathbb{X})^{\mathbb{N}}$ of cardinality $m = \sum_{i=1}^{k} r_i$ is a syntactic cancellative explanation of length m with n repetitions of the pair (x, y) if, for each point of view $i \in N$, the multisets $\langle a_i^{(1)} : r_1, \ldots, a_i^{(k)} : r_k, y_i : n \rangle$ and $\langle b_i^{(1)} : r_1, \ldots, b_i^{(k)} : r_k, x_i : n \rangle$ are equal.*

This definition is illustrated in Ex. 4.
Validation. Checking if a given tuple of pairs of alternatives is an argument of a given pair of alternatives with a given number of repetitions can be performed in $\mathcal{O}(|N| \cdot k \ln k)$, where k is the cardinality of the support set of the explanation.

2.5 The Elliptic Cancellative Argument Scheme

In this section, we propose to streamline the syntactic cancellative argument scheme by omitting the dominance statements. As the resulting scheme is based on an omission (an *ellipsis*), we dub it the *elliptic cancellative scheme*.

Definition 4 (Elliptic cancellative explanation scheme). *Given a dominance relation $\mathcal{D}$, we say the multiset of pairs of alternatives $\langle (a^{(1)}, b^{(1)}) : r_1, \ldots, (a^{(k)}, b^{(k)}) : r_k \rangle \in (\mathbb{X} \times \mathbb{X})^{\mathbb{N}}$ of cardinality $m = \sum_{i=1}^{k} r_i$ is a syntactic cancellative explanation of length m with n repetitions of the pair (x, y) if there exists a multiset of cardinality m' of dominance statements $\langle (c^{(1)}, d^{(1)}) : r_1', \ldots, (c^{(k')}, d^{(k')}) : r_{k'}' \rangle \in \mathcal{D}^{\mathbb{N}}$ such that $\langle (a^{(1)}, b^{(1)}) : r_1, \ldots, (a^{(k)}, b^{(k)}) : r_k \rangle \cup \langle (c^{(1)}, d^{(1)}) : r_1', \ldots, (c^{(k')}, d^{(k')}) : r_{k'}' \rangle$ is a syntactic cancellative explanation of length $m + m'$ with n repetitions of the pair (x, y).*

Example 5. *(Ex. 4 continued) The syntactic cancellative explanation of Ex. 4 can be simplified by removing the last statement d, yielding:*

As:	4*	no	~~15 min~~	180 \$	π_1	2*	~~yes~~	45 min	~~50 \$~~
		~~no~~	45 min	~~50 \$~~	π_2		yes	~~15 min~~	100 \$
So,	4*	no	45 min	180 \$		2*	yes	45 min	100 \$

Validation. It is a little more subtle to check the validity of an elliptic cancellative argument scheme than of a syntactic one. Indeed, when considering the point of view $i \in N$ and comparing the two multisets $L_i := \langle a_i^{(1)} : r_1, \ldots, a_i^{(k)} : r_k, y_i : n \rangle$ and $R_i := \langle b_i^{(1)} : r_1, \ldots, b_i^{(k)} : r_k, x_i : n \rangle$ there are missing elements corresponding to the implicit dominance relations that are not mentioned. Adding these missing dominance relations would have added "good" elements in L_i and "bad" elements in R_i - yielding two lexicographically equivalent vectors. As this is not the case, R_i contains better elements than L_i in the lexicographic sense. In Ex. 5, we obtain $L_{\Psi} = \langle \text{yes} : 1, \text{no} : 2 \rangle$ and $R_{\Psi} = \langle \text{yes} : 2, \text{no} : 1 \rangle$, so that the previous dominance is verified (as R_{Ψ} contains

more "yes" values than L_{Ψ}). Hence the validation of an elliptic cancellative argument scheme simply consists in ordering each L_i and R_i, and checking that, for every j, the jth best elements in R_i is not lesser than the jth best elements in L_i w.r.t. to the order relation $\succsim_i$. Thus, the validation can also be performed in $\mathcal{O}(|N| \cdot k \ln k)$.

3 The Inferred Preference Structure

In this section, we are interested in the description and the computation of the binary relation over alternatives potentially obtained by applying the reasoning engine to the facts of the preference information.

Definition 5. *Given preference information $\mathcal{P}$ and a dominance relation $\mathcal{D}$, we denote $\mathcal{N}_{\mathcal{P},\mathcal{D}}$ the set of pairs of alternatives for which there is a syntactic cancellative explanation of any length with pairs of alternatives in $\mathcal{P} \cup \mathcal{D}$.*

3.1 Inference as Closure

We note that $\mathcal{N}_{\bullet}$ is a closure operator: if new preference statements $\mathcal{N}_{\mathcal{P},\mathcal{D}}$ can be inferred from $\mathcal{P}$ and $\mathcal{D}$, adding them to the knowledge base would not yield additional inference.

Lemma 1.
$$\mathcal{N}_{\mathcal{P},\mathcal{D}} = \mathcal{N}_{\mathcal{N}_{\mathcal{P},\mathcal{D}},\mathcal{D}}$$

Sketch of proof. The inclusion $\mathcal{N}_{\mathcal{P},\mathcal{D}} \subset \mathcal{N}_{\mathcal{N}_{\mathcal{P},\mathcal{D}},\mathcal{D}}$ is a consequence of the fact that $\langle s : 1 \rangle$ is an explanation of s for any statement in $\mathcal{P}$. As for $\mathcal{N}_{\mathcal{P},\mathcal{D}} \supset \mathcal{N}_{\mathcal{N}_{\mathcal{P},\mathcal{D}},\mathcal{D}}$, let $(X, Y) \in \mathcal{N}_{\mathcal{N}_{\mathcal{P},\mathcal{D}},\mathcal{D}}$. There is an syntactic explanation of length m with n repetitions of the pair (X, Y), say $\langle (x^{(1)}, y^{(1)}) : r_1, \ldots, (x^{(K)}, y^{(K)}) : r_K \rangle \in (\mathbb{X} \times \mathbb{X})^{\mathbb{N}}$ where each pair $(x^{(k)}, y^{(k)})$ is in $\mathcal{N}_{\mathcal{P},\mathcal{D}}$, and is therefore supported by an explanation E_k of length m_k with n_k repetitions, with statements in $\mathcal{P} \cup \mathcal{D}$. We claim the tuple obtained by concatenating each explanation E_k repeated $\prod_{k' \in [m], k' \neq k} n_{k'}$ is an explanation with $\prod_{k \in [m]} n_k$ repetitions of the pair (X, Y), with statements in $\mathcal{P} \cup \mathcal{D}$. $\square$

3.2 A Detour via Model-Based Inference

In order to state the main result of this paper, we need to recall the basic principles of *model-based inference*. The goal of inference is extend some (limited) preference information to a richer preference relation $\mathcal{R}$, with 'good' properties, such as $\mathcal{R}$ being a reflexive and, transitive binary relation over $\mathbb{X}$, and maybe complete.

When preference information is given as $\mathcal{P} \cup \mathcal{D}$, where $\mathcal{P}$ is the explicit part, given in so-called *holistic* form, i.e. $\mathcal{P} \subset \mathbb{X}^2$ is a set of reference pairwise statements, and $\mathcal{D}$ is the dominance relation stemming from the ordering of the features, the relation $\mathcal{R}$ is said to be *consistent* when $\mathcal{P} \cup \mathcal{D} \subset \mathcal{R}$.

In order to describe $\mathcal{R}$, which is potentially a very complicated combinatorial object, in a simple language, it is customary to rely on some kind of parameterization of the target set. For instance, numeric models [Jacquet-Lagrèze and Siskos, 1982] in the field of multiple criteria decision making, or graphical languages [Wilson, 2009; Amor *et al.*, 2016] from KR. A popular paradigm consists in considering *value-based*

preferences, where the target preference relation is parameterized by a numeric scoring function $u : \mathbb{X} \to \mathbb{R}$, so that $x\mathcal{R}_u y \iff u(x) \geq u(y)$. (this assumption is made without loss of generality as soon as $\mathcal{R}$ is assumed to be transitive and complete). The target set is still very large and complex, and a common additional assumption is to restrict the scoring function to be *additive* w.r.t. the features, i.e. there is a decomposition such that $\forall x \in \mathbb{X}, u(x) = \sum_{i \in N} \omega_i(x_i)$.

Definition 6 (preferences based on additive value). *The parameter set of the additive value model is $\Omega_\Sigma := \prod_{i \in N} \mathbb{R}^{\mathbb{X}_i}$, and for a given value $\omega := \langle \omega_i \rangle_{i \in N}$ of the parameter, the corresponding preference relation $\mathcal{R}_{\sum \omega} \subset \mathbb{X}^2$ is defined by:*

$$\forall x, y \in \mathbb{X}, x\mathcal{R}_{\sum \omega} y \iff \sum_{i \in N} \omega_i(x_i) \geq \sum_{i \in N} \omega_i(y_i).$$

For such an additive value, that we denote $\mathcal{R}_{\sum \omega}$, the condition $\mathcal{D} \subset \mathcal{R}_{\sum \omega}$ translates to the following monotonicity conditions: for all features i, the function $\omega_i : (\mathbb{X}_i, \succsim_i) \to (\mathbb{R}, \geq)$ is nondecreasing.

The most prevalent approach in model-based inference consists in determining the most adequate value of the parameter in the sense of some loss function $\mathcal{L}$: $\omega^\star = \mathrm{argmin}_{\Omega_\Sigma} \mathcal{L}$, and returning the corresponding preference relation $\mathcal{R}_{\sum \omega^\star}$. Meanwhile, the *robust* approach consists in considering the intersection of all the consistent preference relations, assuming it is not empty:

$$\mathcal{R}^\star_{\Omega_\Sigma} := \bigcap_{\omega \in \Omega_\Sigma : (\mathcal{P} \cup \mathcal{D}) \subset \mathcal{R}_{\sum \omega}} \mathcal{R}_{\sum \omega}.$$

The robust approach yields the *version space* [Mitchell, 1982] of the model. Equivalently, it can be understood as assuming that the preference information $\mathcal{P}$ is *incomplete* (as there might be several value of the parameter that are consistent with it), and drawing skeptical conclusions with respect to all the possible completions.

3.3 Cancellative-Powered Deductions are Robust Inferences under Additive Values

We are now able to state an important result concerning the preference structure $\mathcal{N}_{\mathcal{P},\mathcal{D}}$.

Theorem 1.
$$\mathcal{N}_{\mathcal{P},\mathcal{D}} = \mathcal{R}^\star_{\Omega_\Sigma}$$

The inferred preference structure is exactly the necessary preference relation under the assumption of an additive value model. This result has an important corollary concerning the inferred relation:

Corollary 1 (Properties of the inferred structure). *$\mathcal{N}_{\mathcal{P},\mathcal{D}}$ is a transitive and reflexive binary relation.*

The proof of Th. 1 relies on the fact that, under the assumption of an additive value model, a preference statement can be represented by a linear form operating over the vector space Ω_Σ.

Definition 7. *Given some preference information $\mathcal{P} \subset \mathbb{X}^2$, alternatives $x, y \in \mathbb{X}$, and a point of view $i \in N$, for any value $x_i \in \mathbb{X}_i$, let $\epsilon_{i,x_i} : \mathbb{R}^{\mathbb{X}_i} \to \mathbb{R}, \omega_i \mapsto \omega_i(x_i)$, and*

$\phi_{(x,y)} = \sum_{i \in N} \epsilon_{i,x_i} - \epsilon_{i,y_i}$ *a linear form over $\mathbb{R}^{\mathbb{X}}$. Also, let $\widehat{\mathbb{X}}_i := \{t \in \mathbb{X}_i : \exists (a,b) \in \mathcal{P}, t = a_i \text{ or } t = b_i\} \cup \{x_i\} \cup \{y_i\}$ and $\widehat{X} := \prod_{i \in N} \widehat{\mathbb{X}}_i$.*

Lemma 2.

$$(x,y) \in \mathcal{R}_{\sum \omega} \iff \phi_{(x,y)}(\omega) \geq 0$$

Proof of $\mathcal{N}_{\mathcal{P},\mathcal{D}} \subset \mathcal{R}^\star_{\Omega_\Sigma}$.
Let $(x,y) \in \mathcal{N}_{\mathcal{P},\mathcal{D}}$. By definition, there is a syntactic cancellative explanation of length m with n repetitions of the pair (x,y), say $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathcal{P} \cup \mathcal{D})^m$. Therefore, for each point of view $i \in N$, $(y_i, \ldots, y_i, a_i^{(1)}, \ldots, a_i^{(m)})$ is a permutation of $(x_i, \ldots, x_i, b_i^{(1)}, \ldots, b_i^{(m)})$. In particular, for any parameter $\omega \in \Omega_\Sigma$, $n\omega_i(y_i) + \omega_i(a_i^{(1)}) + \cdots + \omega_i(a_i^{(1)}) = n\omega_i(x_i) + \omega_i(b_i^{(1)}) + \cdots + \omega_i(b_i^{(1)})$, so $n\phi_{(x,y)}(\omega) = \sum_{j=1}^m \phi_{(a^{(j)},b^{(j)})}(\omega)$. Now, if ω is consistent, $(\mathcal{P} \cup \mathcal{D}) \subset \mathcal{R}_\omega$ and $\phi_{(a^{(j)},b^{(j)})}(\omega) \geq 0$. Thus $n\phi_{(x,y)}(\omega)$ is nonegative as the sum of m nonnegative terms, and x is necessarily preferred to y under the assumption of an additive value model. $\square$

Proof of $\mathcal{N}_{\mathcal{P},\mathcal{D}} \supset \mathcal{R}^\star_{\Omega_\Sigma}$.
Let $(x,y) \in \mathcal{R}^\star_{\Omega_\Sigma}$. For any parameter $\omega \in \Omega_\Sigma$ such that $\forall s \in (\mathcal{P} \cup \mathcal{D}), \phi_s(\omega) \geq 0$, $\phi_{(x,y)}(\omega) \geq 0$. This property concerns linear forms in $\mathbb{R}^{\mathbb{X}}$, which is a vector space of infinite dimension, but also holds in $\mathbb{R}^{\widehat{\mathbb{X}}}$. Indeed, $\widehat{\mathbb{X}} \subset \mathbb{X}$, is of finite dimension, and any additive parameter function $\widehat{\omega} : \widehat{\mathbb{X}} \to \mathbb{R}$ such that $(\mathcal{P} \cup \mathcal{D}) \cap \widehat{\mathbb{X}}^2 \subset \mathcal{R}_{\widehat{\omega}}$ can be extended into an additive function $\omega : \mathbb{X} \to \mathbb{R}$ describing a consistent relation $\mathcal{R}_{\sum \omega}$. By Farkas' lemma, the linear form $\phi_{(x,y)}$ is a conical combination of the $\langle \phi_s \rangle_{s \in (\mathcal{P} \cup \mathcal{D}) \cap \widehat{\mathbb{X}}^2}$. As the coefficients of all these linear forms are integers—they are, indeed, in $\{-1, 0, 1\}$—the coefficients of the conical combinations can be chosen rational, and by multiplying by the lesser common multiple of their denominators, yield an identity: $n \, \phi_{(x,y)} = \sum_{s \in (\mathcal{P} \cup \mathcal{D}) \cap \widehat{\mathbb{X}}^2} m_s \, \phi_s$, with a positive integer n and nonnegative integer coefficients $\langle m_s \rangle_{s \in (\mathcal{P} \cup \mathcal{D}) \cap \widehat{\mathbb{X}}^2}$. We claim the tuple $\langle s : m_s \rangle_{s \in (\mathcal{P} \cup \mathcal{D}) \cap \widehat{\mathbb{X}}^2}$ is a syntactic cancellative explanation of length n of the pair (x,y), with statements in $\mathcal{P} \cup \mathcal{D}$, thus $(x,y) \in \mathcal{N}_{\mathcal{P} \cup \mathcal{D}}$. $\square$

3.4 Efficient Inference Procedures

The necessary relation assuming an additive value model $\mathcal{R}^\star_{\Omega_\Sigma}$ is defined and studied by [Greco *et al.*, 2008]. In particular, Greco et al. propose a linear program permitting to solve the decision problem corresponding to our research question concerning inference: given a pair of alternative, decide if it is in the inferred preference relation. This linear program is expressed in the primal space $\mathbb{R}^{\widehat{X}}$ of the values $\omega_i(x_i)$ given to each relevant value of the attributes. These values are of little interest concerning our cancellative argument schemes, so we propose to formulate the dual problem.

Corollary 2 (Polytime inference via conical decomposition). *For all pairs of alternatives $(x,y) \in \mathbb{X}^2$, $(x,y) \in \mathcal{N}_{\mathcal{P},\mathcal{D}}$ if,*

and only if, the following linear program is feasible:
find nonnegative real numbers $\langle\lambda_s\rangle_{s\in(\mathcal{P}\cup\mathcal{D})\cap\widehat{X}^2}$ *such that*

$$\phi_{(x,y)}=\sum_{s\in(\mathcal{P}\cup\mathcal{D})\cap\widehat{X}^2}\lambda_s\phi_s.$$

Would the decision variables λ be integers, they could directly be interpreted as a multiset $\langle s:\lambda_s\rangle$ serving as a syntactic cancellative explanation for (x,y). As the elliptic scheme tells us that the coefficients corresponding to the dominance statements are eventually irrelevant, this formulation ought to be further streamlined. Unfortunately, the conical span of the dominance statements is not easy to characterize in the dual base (ϵ_{i,x_i}). This obstacle can be lifted by representing the preference statements in an alternative decomposition, that focuses on differences of values, rather than values.

Definition 8. *Given a finite binary relation $A\subset\mathbb{X}^2$, for all points of view $i\in N$ we denote $\{\widehat{x}_{i,1}\precsim\widehat{x}_{i,2}\precsim\cdots\precsim\widehat{x}_{i,|\widehat{\mathbb{X}}_i|}\}=\widehat{\mathbb{X}}_i$. For any integer $k,1\leq k<|\widehat{\mathbb{X}}_i|$, let $\delta_{\widehat{\mathbb{X}},i,k}:=\epsilon_{i,\widehat{x}_{i,k+1}}-\epsilon_{i,\widehat{x}_{i,k}}.$*

Lemma 3. *For any statement $(x,y)\in A$ and any point of view $i\in N$,*

$$\epsilon_{i,x_i}-\epsilon_{i,y_i}=\begin{cases}0 & \text{if }x_i\sim_i y_i\\ \sum_{k:x_i\leq\widehat{x}_{i,k}<y_i}(-1)\cdot\delta_{\widehat{\mathbb{X}},i,k}, & \text{if }x_i\prec_i y_i\\ \sum_{k:y_i\leq\widehat{x}_{i,k}<x_i}(+1)\cdot\delta_{\widehat{\mathbb{X}},i,k}, & \text{if }x_i\succ_i y_i\end{cases}$$

This lemma has two important consequences:

i) Corollary 2 can be expressed in terms of $\langle\delta_{\widehat{X}^2}\rangle$ rather than $\langle\epsilon\rangle$; and

ii) dominance statements in A are exactly the conical span of the $\langle\delta_A\rangle$.

This leads to a leaner reformulation of the inference problem.

Definition 9. *For all $x,y\in\mathbb{X}$, $i\in N$ and $k\in\mathbb{N}:1\leq k<|\widehat{\mathbb{X}}_i|$, let*

$$\varphi_{(x,y)}^{(i,k)}:=\begin{cases}-1, & \text{if }x_i\precsim_i\widehat{x}_{i,k}\prec_i y_i;\\ 0, & \text{if }x_i\sim_i y_i;\text{ or}\\ +1, & \text{if }y_i\precsim_i\widehat{x}_{i,k}\prec_i x_i.\end{cases}$$

Theorem 2 (Inference via LP). *For all pairs of alternatives $(x,y)\in\mathbb{X}^2$, $(x,y)\in\mathcal{N}_{\mathcal{P},\mathcal{D}}$ if, and only if, the following linear program is feasible:*
find nonnegative real numbers $\langle\lambda_s\rangle_{s\in\mathcal{P}}$ such that the inequality $\varphi_{(x,y)}^{(i,k)}\geq\sum_{s\in\mathcal{P}}\lambda_s\varphi_s^{(i,k)}$ holds for every indices $i\in N$ and $1\leq k<|\widehat{\mathbb{X}}_i|$.

3.5 Repetition of the Conclusion

The presence of repetition of the conclusion makes the explanation scheme cumbersome. One may wonder whether it is possible to get rid of the repetition of the conclusion.

Theorem 3. *It is not possible to cover all possible inferences obtained by the robust additive model by restricting the cancellation explanation schema with $n=1$ repetition.*

Proof. We provide a counter-example with $|N|=6$ features, $\mathbb{X}=\{0,1\}^6$ and $1\succsim_i 0$ (statement d_i) for all $i\in N$. The preference information is:

$$\pi_1:\big((0,0,1,\cdot,\cdot,\cdot),(1,1,0,\cdot,\cdot,\cdot)\big)$$
$$\pi_2:\big((0,\cdot,\cdot,0,1,\cdot),(1,\cdot,\cdot,1,0,\cdot)\big)$$
$$\pi_3:\big((\cdot,0,\cdot,\cdot,1,0),(\cdot,1,\cdot,\cdot,0,1)\big)$$
$$\pi_4:\big((\cdot,\cdot,1,0,\cdot,0),(\cdot,\cdot,0,1,\cdot,1)\big)$$

We can infer from the preference information that $(0,0,1,0,1,0)$ is preferred to $(1,1,0,1,0,1)$ (statement π_C). One can readily see that $\pi_1+\pi_2+\pi_3+\pi_4=2\,\pi_C$.

Assuming by contradiction that there exist Farkas coefficients with coefficient 1 associated to π_C: $\pi_C=\sum_{i=1}^6\lambda_i\pi_i+\sum_{i=1}^6\mu_i\,d_i$, where $\lambda_i,\mu_i\in\mathbb{N}$ leads to an infeasible linear system.
$\qquad\square$

One could also want to trim down the potential complexity of the explanations by limiting the number of premises. Unfortunately, this might lead to loss of transitivity for the inferred relation.

4 Explanations for Valid Preference Statements

It seems reasonable to believe that an explanation is easier to process by a cognitive agent—'simpler'—when it is short. In the case of cancellative explanations, the actual cognitive burden mainly comes from three factors: the number of points of view $|N|$, that we consider as mostly exogenous; the length m of the premises; and the number n of repetitions of the conclusion. Without any experimental evidence, we consider the problem of finding an explanation for a given pair $(x,y)\in\mathcal{N}_\mathcal{P}$ which is as simple as possible as a bi-objective integer linear minimization problem:

$$\min_{n,m\in\mathbb{N}^\star}(n,m)\quad\text{such that}\begin{cases}n\,\varphi_{(x,y)}\geq\displaystyle\sum_{\pi\in\mathcal{P}}\ell_\pi\varphi_\pi;\quad\text{and}\\ m\geq\displaystyle\sum_{\pi\in\mathcal{P}}\ell_\pi.\end{cases}$$
$$(1)$$

Integer linear programs offer a powerful language permitting to describe difficult combinatorial problems. These formulations can be given wholesale to dedicated solvers, that eschews the need for developing a dedicated piece of software and benefits from state-of-the-art refinements in the solving of such problems. Nevertheless, it would be unwise to delegate the search for a short explanation of a given pair of alternatives to such a solver, if this search were not, intrinsically, a difficult combinatorial problem. The following theorem addresses this issue.

Theorem 4. *The problem of deciding, for a given input $(x,y,n,m)\in\mathbb{X}\times\mathbb{X}\times\mathbb{N}^\star\times\mathbb{N}^\star$ if there is an elliptic cancellative explanation of the pair (x,y) of length at most m with at most n repetitions is NP-complete. This remains true even if the number n of repetitions is set to one.*

Proof. Membership to NP is ensured, as checking the validity of an elliptic scheme is polyomial in the number of distinct premises, which is upper bounded by the cardinality of $\mathcal{P}$.

Hardness can be established e.g. by reduction from VER-TEX COVER [Karp, 1972]. Formally, a vertex cover V' of an undirected graph $G = (V, E)$ is a subset of V such that $uv \in E \Rightarrow u \in V' \vee v \in V'$, that is to say it is a set of vertices V' where every edge has at least one endpoint in the vertex cover V'. The VERTEX COVER problem consists in, given an instance (G, k) where $G = (V, E)$ is a graph and k a positive integer, to *decide* whether G has a vertex cover of size at most k, or not. Given an instance of VERTEX COVER, we map it to a gadget instance of our problem:

- the set of points of view is $N = V \cup E$;

- an alternative is a subset of N;

- each point of view is evaluated on a binary scale, with presence preferred to absence;

- the preference information contains all statements of the form $(\{(u, v)\}, \{u, v\})$—any edge is preferred to the set of its endpoints—for all edges $(u, v) \in E$.

Any elliptic cancellative explanation without repetition of the pair (E, V)—the pros are the edges, the cons are the vertices—of length k is a subset of E that forms a vertex cover of size k of of the graph G, and reciprocally. $\qquad\square$

5 Discussion and Perspectives

In the current quest for "explainable A.I.", the additive value (i.e. linear) model might be seen as occupying the very end of the spectrum –an obviously interpretable model [Ribeiro *et al.*, 2016]. Even though recent advances have been made towards providing "simpler" models, e.g. [Ustun and Rudin, 2016], most of these approaches ignore the perspective of the decision maker [Miller, 2019], and the need to provide her with a way of challenging the decision [Kroll *et al.*, 2017].

Several works have explored the interplay between argumentation and decision aiding. In [Amgoud and Prade, 2009], argumentation is used as a mean make a decision and justify it, while in [Zhong *et al.*, 2019], it is shown that the outcomes of a simple decision model are similar to the extension of the corresponding argumentation framework.

Here, the preference information is considered exogenous. It might have been obtained through dialog, by considering domain knowledge—reference cases, jurisprudence, or inferred by some means—learning from similar situations, or previous interactions with the user.

5.1 Contributions

We introduced the notion of *cancellative explanations*, based on the accrual of premises to obtain a conclusion. We studied this explanative framework in the light of the principles stated in introduction. This contrasts with approaches in decision theory [Fishburn, 1970; Gonzales, 2000], where cancellation is seen as a property of the preference relation, not a mean to infer new preference statements and justify them. Our main contributions are as follows:

Completeness. Every preference statement that can be skeptically inferred from the preference information and the way of reasoning corresponding to the additive value model is supported by a cancellative explanation.

Soundness. Every preference statement that is supported by a cancellative explanation can be skeptically inferred from the preference information and the way of reasoning corresponding to the additive value model;

Simplicity. We provided several ways of presenting cancellative explanations, in the form of tables, diagrams, or argument schemes, and proposed to ground them on a syntactic check, or alternatively to keep implicit the information tied to dominance, which can easily be restored by the recipient, in the spirit of *enthymemes*. We provided formalizations that lend themselves to an efficient implementation. We proposed an intuitive partial ordering of explanations according to their alleged complexity, and formulated the problem of finding explanations as simple as possible.

Computation. Remarkably, while adjudicating necessary preference is a polynomial problem, explaining it concisely is NP-complete.

5.2 Perspectives

Providing an argument scheme along with the result of a comparative statement opens the possibility to discuss or challenge this result. This is made possible through what is called critical questions [Walton, 1996], a tool associated with argument schemes representing attacks or criticisms that, if not answered adequately, falsify the argument fitting the scheme.

In our setting, the criticism may point out (implicitly or explicitly) elements perceived as missing or wrong in the reasoning steps. Indeed, for instance, the decision maker (DM) may challenge the fact that a preference between two alternatives is not the right one. The consequence is that either it is possible to derive a new conclusion with this new information, or the DM's statements express conflicting preferences. Thus, the challenge of finding a principled way to deal with inconsistency in an accountable manner, needs to be addressed. Several promising approaches have been proposed: considering maximally consistent subsets of statements [Mousseau *et al.*, 2003]; relaxing the aggregation model until a model sufficiently expressive to accommodate for the preference information is found [Ouerdane, 2011; Greco *et al.*, 2014]; or using a numerical estimation of inconsistency such as a belief function [Destercke, 2018].

Another situation is that the DM's reasoning is incompatible with the principles and properties underlying the preference model. For instance, expressing a preference dependency may defeat the fundamental feature (ceteris paribus) of an additive model [Fisher, 1892]. In this situation, relaxing the preference model could be a solution [Ouerdane, 2011]. Many models account for interactions between the influence of the points of view, such as Generalized additive models (GAI) [Fishburn, 1967]. An underlying question that has been less investigated (for notable exceptions, see e.g. [Labreuche, 2011] and [Cailloux and Endriss, 2014]), and remains difficult [Procaccia, 2019], is the question of the accountability of recommendations based on an induced model.

Acknowledgements

References

[Amgoud and Prade, 2009] Leila Amgoud and Henri Prade. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173:413–436, 2009.

[Amor *et al.*, 2016] Nahla Ben Amor, Didier Dubois, Hela Gouider, and Henri Prade. Graphical models for preference representation: An overview. In *Proceedings of the 10th International Conference on SUM*, pages 96–111, 2016.

[Bouyssou *et al.*, 2009] Denis Bouyssou, Didier Dubois, Marc Pirlot, and Henri Prade. *Decision Making Process: Concepts and Methods*. Wiley ISTE, 2009.

[Cailloux and Endriss, 2014] Olivier Cailloux and Ulle Endriss. Eliciting a suitable voting rule via examples. In *Proceedings of the 21st ECAI'14* , pages 183–188, 2014.

[Destercke, 2018] Sébastien Destercke. A generic framework to include belief functions in preference handling and multi-criteria decision. *International Journal of Approximate Reasoning*, 98:62 – 77, 2018.

[Dubois *et al.*, 2008] Didier Dubois, Hélène Fargier, and Jean-Fraccois Bonnefon. On the qualitative comparison of decisions having positive and negative features. *J. Artif. Intell. Res.*, 32:385–417, 2008.

[Fishburn, 1967] Peter C. Fishburn. Interdependence and additivity in multivariate, unidimensional expected utility theory. *International Economic Review*, 8(3):335–342, 1967.

[Fishburn, 1970] Peter C. Fishburn. *Utility Theory for Decision Making*. J. Wiley & Sons, 1970.

[Fisher, 1892] Irving Fisher. *Mathematical investigations in the theory of value and prices, and appreciation and interest*. 1892.

[Gonzales, 2000] Christophe Gonzales. Two factor conjoint measurement with one solvable component. *Journal of Mathematical Psychology*, 44(2):285–309, 2000.

[Greco *et al.*, 2008] Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *EJOR*, 191(2):416–436, 2008.

[Greco *et al.*, 2014] Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Robust ordinal regression for value functions handling interacting criteria. *EJOR*, 239(3):711–730, 2014.

[Hammond *et al.*, 1998] John Hammond, Ralph Keeney, and Howard Raiffa. Even Swaps: a rational method for making trade-offs. *Harvard Business Review*, pages 137–149, 1998.

[Jacquet-Lagrèze and Siskos, 1982] Eric Jacquet-Lagrèze and Yanis Siskos. Assessing a set of additive utility functions for multicriteria decision making: the UTA method. *EJOR*, 10:151–164, 1982.

[Karp, 1972] Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations*, pages 85–103. 1972.

[Krantz *et al.*, 1971] David H. Krantz, Duncan R. Luce, Patrick Suppes, and Amos Tversky. *Foundations of measurement*, volume 1: Additive and Polynomial Representations. Academic Press, 1971.

[Kroll *et al.*, 2017] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2017.

[Labreuche, 2011] Christophe Labreuche. A general framework for explaining the results of a multi-attribute preference model. *AIJ*, 175:1410–1448, 2011.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

[Mitchell, 1982] Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[Mousseau *et al.*, 2003] Vincent Mousseau, Luis C. Dias, José Figueira, Carlos Gomes, and João N. Clímaco. Resolving inconsistencies among constraints on the parameters of an MCDA model. *EJOR*, 147(1):72–93, 2003.

[Ouerdane, 2011] Wassila Ouerdane. Multiple criteria decision aiding: a dialectical perspective. *4OR*, 9:429–432, 2011.

[Procaccia, 2019] Ariel D. Procaccia. Axioms should explain solutions. In Laslier, Moulin, Sanver, and Zwicker, editors, *Future of Economic Design*. 2019.

[Ribeiro *et al.*, 2016] Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*, pages 1135–1144, 2016.

[Toulmin, 1958] Stephen Toulmin. *The uses of Arguments*. Cambridge University Press, 1958.

[Ustun and Rudin, 2016] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.*, 102(3):349–391, 2016.

[Wakker, 1989] Peter Wakker. *Additive Representations of Preferences: A New Foundation of Decision Analysis*. Theory and Decision Library C. 1989.

[Walton, 1996] Douglas Walton. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996.

[Wilson, 2009] Nic Wilson. Efficient inference for expressive comparative preference languages. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 961–966, 2009.

[Zhong *et al.*, 2019] Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications*, 117:42–61, 2019.

A Model for Accountable Ordinal Sorting

Khaled Belahcene[1]**, Christophe Labreuche**[2]**, Nicolas Maudet**[3]**, Vincent Mousseau**[1]**, Wassila Ouerdane** [1]

[1] LGI, CentraleSupélec, Université Paris-Saclay, Châtenay Malabry, France.
[2]Thales Research & Technology, 91767 Palaiseau Cedex, France.
[3]Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 75005 Paris.
khaled.belahcene, vincent.mousseau, wassila.ouerdane@centralesupelec.fr
christophe.labreuche@thalesgroup.com   nicolas.maudet@lip6.fr

Abstract

We address the problem of multicriteria ordinal sorting through the lens of *accountability*, *i.e.* the ability of a human decision-maker to own a recommendation made by the system. We put forward a number of model features that would favor the capability to support the recommendation with a convincing explanation. To account for that, we design a recommender system implementing and formalizing such features. This system outputs explanations under the form of specific *argument schemes* tailored to represent the specific rules of the model. At the end, we discuss possible and promising argumentative perspectives.

1   Introduction

While algorithmic automated decisions or recommendations are nowadays pervasive, there is a growing demand of institutions and citizens to make these recommendations *transparent* and *trustworthy*, while system designers seek *persuasive* recommendations [Tintarev, 2007]. The recent regulation adopted by the European Parliament (known as the General Data Protection Regulation, GDPR) goes further by adding a "right to explanation". According to [Goodman and Flaxman, 2016] *"the GDPR's requirements could require a complete overhaul of standard and widely used algorithmic techniques"*. We interpret this requirement in the strong sense of *accountability*, its litmus test being the ability of the recipient of the recommendation to defend it before other, skeptical, stakeholders of the decision (whereas *trust* requires the recommendation to be consistently accurate, but eventually asks for delegation of the decision to the system; *transparency* simply provides access to the underlying algorithm without concern for technical literacy [Burell, 2016]; and *persuasiveness* is hardly transferable: someone persuaded by a recommendation may not be a good persuader).

Our aim in this paper is thus to build an accountable, ordinal, multicriteria classifier, mapping a *candidate* object to a *recommendation* consisting in one or more categories among a predefined, ordered collection of these. In a multicriteria decision aiding (MCDA) context, the only indisputable relation between objects is the Pareto dominance, occurring when an object outperforms another on all criteria. As the situation

is seldom so clear, the rules permitting the comparison of objects need to be enriched, taking into account the knowledge and values of the decision-maker, collected under the label *preference information*, which is also considered as an input of the classifier. We also consider an additional output, an *explanation* aimed at the decision-maker, supporting the recommendation and enabling the accountability sought for. In order to reach this goal of accountability, we make two important assumptions about the recommender system. These *design principles* are as follows:

No jargon. A first step in a MCDA process is to collect decision-maker's preferences information. In order to accurately represent the specific decision process, we opt for an indirect elicitation [Dias *et al.*, 2002]: the decision-maker is never asked any questions about artifacts of the model (e.g. weights). Instead she should express preferences directly in the language of the actual decision situation, i.e. providing direct assignments of typical examples, *reference objects*, to categories.

No arbitrariness. MCDA usually proceeds by representing the reasoning of the decision-maker with a formal parametric model, describing a specific stance. The values of the *preference parameters* are often fitted during an elicitation process, up to a certain point. While many methods proceed by picking a specific, so-called *representative* value of the parameters, we opt for a *robust* approach (to the lack of preference information) [Vincke, 1999; Greco *et al.*, 2008], formulating a –possibly partial – recommendation that cannot be refuted by any judgment function consistent with the preference information.

On top of these principles, we make three further assumptions about the MCDA model, proceeding from the willingness to keep the model accessible to human reasoning.

No compensation. This assumption deals with the interpretation of collected data –the evaluation of objects on various criteria. We assume they are always used comparatively, in a purely ordinal manner: on a given criterion, an alternative is either as good as another one, or strictly worse. Hence, only the *set* of criteria for which an alternative is better is important, regardless of the specific values, and being very good on some criterion cannot compensate for low performance on others. This feature enables the algorithm to proceed without performing any algebraic computation, which makes it particularly suited for explanation. It is shared with established

non-compensatory ordinal sorting models used in the field of MCDA (eg. NCS) [Bouyssou and Marchant, 2007]. Moreover, the use of a 2-valued comparison $(\geq, <)$ is similar to [Bouyssou, 1986] rather than [Fishburn, 1976] who proposes a 3-valued one $(<, =, >)$.

No values. At the heart of the recommender system is a *preference structure* encoding the comparison of alternatives. There are two main families of structures: those based on *value* [Keeney and Raiffa, 1976], and those based on *outranking relations* [Roy, 1991]. We opt for the latter, as they eschew the construction of a scoring function. An outranking relation naturally provides four outcomes when comparing two alternatives: preference for the former, for the latter, indifference, or incomparability; also, it does not enforce transitivity of preference.

No frontiers. In MCDA, most classifiers link the preference structure and the recommendation of a class by introducing an explicit frontier between classes, defining the limit of each class (a single value for value-based models, a limiting profile for outranking-based ones, e.g. [Leroy *et al.*, 2011]). We do without this construct, as for instance models based on Logical Analysis of Data (LAD) techniques [Crama *et al.*, 1988] which output classification rules. We shall use simple rules permitting to classify a new object by comparing it to a set of already classified *reference objects* (see Sect.2.3).

The general philosophy of these principles must be clear to the reader: accountability should exclude in principle the use of any model artifact that the decision-maker may not properly handle, but at the same time provide enough understanding of the model so as to allow the decision-maker to defend the recommendation *as if it was her own*. Following this, our approach is to enforce these principles by design, and to investigate how far we can get with the resulting sorting model. This approach differs from the recent work of [Ribeiro *et al.*, 2016] which adopts a model-agnostic approach, and builds explanations adapted to virtually any classifier. They obtain extremely promising results in terms of trust. As expected, the explanation cannot be fully faithful to the model (they are "locally" faithful though). It also differs from [Datta *et al.*, 2016] which seeks to extract how influential are input parameters, but keeping a black-box access to the model. While for the trust requirement these approaches are sufficient, our notion of accountability requires to get to grips with the model.

The rest of this paper is as follows. We propose a model implementing and formalizing the different principles, decomposing it in a learning phase (Section 2) and a recommendation phase (Section 3). We provide formal explanations of the recommendation in most cases, in the form of *argument schemes* tailored to represent the specific rules of the model. Section 4 introduces some insights on the description of the sorting problem through an argumentation system. Section 5 concludes the paper, by putting its findings into perspective.

2 Formal Description

In this section, we define a recommender system following the design principles and assumptions, and describe some of its properties.

2.1 The Recommender System

We consider a multicriteria ordinal sorting problem : a collection of objects are evaluated on a set of criteria N. We note $\mathbb{B} := \{0, 1\}$, so that elements of $\mathbb{B}^N$ are at the same time vectors with binary coordinates, and subsets of N, partially ordered by inclusion. The maximal element of $\mathbb{B}^N$ is the unanimous coalition N, also denoted $(1, \dots, 1)$. The minimal element of $\mathbb{B}^N$ is the empty coalition $\emptyset$, also denoted $(0, \dots, 0)$. Each criterion $i \in N$ maps an object to a performance value in a totally ordered set $\mathbb{X}_i$, the higher the better. Consequently, each object is described by a performance vector in the partially ordered set $\mathbb{X} = \prod_{i \in N} \mathbb{X}_i$. The objects are to be assigned to some class chosen among an ordered set $\mathbb{K} = \{k_1 \prec \cdots \prec k_p\}$, so that assignment to a class with a high index is desirable.

Formally, let us describe the recommender system as a function mapping a pair $\langle z, \mathcal{P} \rangle$ to a pair $\langle K, \mathcal{E} \rangle$, where:

- The object $z \in \mathbb{X}$ is a *candidate* for sorting;

- $\mathcal{P}$ denotes *preference information* collected from the decision-maker consisting of typical classification examples, a collection of *reference objects* $\mathbb{X}^\star \subset \mathbb{X}$, and their assigned categories $Class : \mathbb{X}^\star \to \mathbb{K}$. For syntactic reasons, we represent it by a set of object-assignment pairs $\mathcal{P} \subset \mathbb{X} \times \mathbb{K}$.

$$\mathcal{P} := \bigcup_{x^\star \in \mathbb{X}^\star} (x^\star, Class(x^\star))$$

- $K \subset \mathbb{K}$ is the *recommendation*, concerning the classes that could be assigned to the candidate (see Sect. 3);

- $\mathcal{E}$ is an *explanation* yet unspecified, supporting the recommendation K (see for instance [Labreuche *et al.*, 2012; Belahcene *et al.*, 2017]), and addressed by Sect. 3.

Example 1. *Objects are evaluated according to four criteria a, b, c, d (higher is better). Six reference objects: $\mathbb{X}^\star := \{A_1, A_2, B_1, B_2, C_1, C_2\}$, described by the performance table below, are assigned to three classes: $\mathbb{K} := \{\star \prec \star\star \prec \star\star\star\}$ and make up the preference information $\mathcal{P}$. We consider two candidates: X, Y and try to assign them to some possible classes.*

Object	a	b	c	d	Assignment
A_1	3	3	2.5	0	$\star\star\star$
A_2	3	2	2.1	1	$\star\star\star$
B_1	2	2	1.3	1	$\star\star$
B_2	3	1	3.7	0	$\star\star$
C_1	2	1	1.6	1	$\star$
C_2	1	1	4.1	0	$\star$
X	2	2	1.1	0	?
Y	2	3	1.8	0	?

2.2 The Reasoning of the Decision-Maker

A non-compensatory outranking relation can be represented by a Boolean composite function:

$$\forall x, y \in \mathbb{X}, \ x S_\phi y \iff \phi \circ O_N(x, y) = 1$$

where the *observation function* O_N maps a pair of objects to its *concordance set*, and the consistent judgment of the decision-maker, based on these concordance sets, is represented by the *judgment function* ϕ mapping a concordance set to a truth value.

$$O_N : \quad \mathbb{X} \times \mathbb{X} \quad \rightarrow \quad \mathbb{B}^N$$
$$(x, y) \quad \mapsto \quad \{i \in N : x_i \geq y_i\}$$

Antecedents of 1 by ϕ, called *true points* in the language of the LAD [Crama *et al.*, 1988], represent *sufficient coalitions of criteria*, while antecedents of 0 by ϕ are *false points* or *insufficient coalitions of criteria*. ϕ is supposed *non-decreasing*, meaning that a superset of a sufficient coalition of criteria is also sufficient, and a subset of an insufficient coalition is also insufficient. Compatibility of the outranking relation S to the Pareto dominance imposes that a unanimous support of criteria is always sufficient, so $\phi(N) = 1$. Conversely, $\phi(\emptyset) = 0$ must hold, so the relation S is not reduced to generalized indifference. Finally, we define the set of any possible judgment function :

$$\phi \in \widehat{\Phi} := \{\phi : \mathbb{B}^N \to \mathbb{B} : \phi \nearrow \text{ and } \phi(N) = 1 \text{ and } \phi(\emptyset) = 0\}$$

2.3 Learning From the Assignment Examples

To assign a new object to a category, we shall use the following classification rules:

- *(R1)* an object cannot outrank any object assigned to a strictly better class;

- *(R2)* an object outranks objects assigned to a strictly worse class;

- *(R3)* objects in the same class can be in any position with respect to outranking.

To account for that, we first denote $\succsim_\mathcal{P}$ the complete pre-order between reference objects induced by $\mathcal{P}$:

$$\begin{cases} x^\star \succsim_\mathcal{P} y^\star & \Longleftrightarrow & Class(x^\star) \succsim Class(y^\star) \\ x^\star \succ_\mathcal{P} y^\star & \Longleftrightarrow & Class(x^\star) \succ Class(y^\star) \\ x^\star \sim_\mathcal{P} y^\star & \Longleftrightarrow & Class(x^\star) = Class(y^\star) \end{cases}$$

We consider the strict enforcement of the model rules for reference objects:

- *(R1)* : $\forall x^\star, y^\star \in \mathbb{X}^\star, x^\star \succ_\mathcal{P} y^\star \Rightarrow Not\ (y^\star S_\phi x^\star)$;

- *(R2)* : $\forall x^\star, y^\star \in \mathbb{X}^\star, x^\star \succ_\mathcal{P} y^\star \Rightarrow x^\star S_\phi y^\star$.

Hence, the assignment of reference objects expressed by $\mathcal{P}$ places upper (by *(R1)*) and lower (by *(R2)*) bounds upon the outranking relation between reference objects. so that:

$$\succ_\mathcal{P} \subseteq S_\phi \cap (\mathbb{X}^\star)^2 \subseteq \succsim_\mathcal{P}$$

These constraints transfer to the judgment functions. Each pair $(x^\star, y^\star)$ is mapped by the observation function O_N to a coalition of criteria. The observed coalitions $O_N(\mathbb{X}^\star \times \mathbb{X}^\star)$ serve as a learning set for the judgment function ϕ. They are sorted between three sets, yielding necessary conditions on ϕ:

- insufficient coalitions $O_N(\prec_\mathcal{P})$ should be mapped to 0;

- sufficient coalitions $O_N(\succ_\mathcal{P})$ should be mapped to 1;

- $O_N(\sim_\mathcal{P})$, which images by ϕ are not constrained.

Consequently, we define the set $\Phi(\mathcal{P})$ of judgment functions compatible to the preference information $\mathcal{P}$:

$$\Phi(\mathcal{P}) := \{\phi \in \widehat{\Phi} \ : \ \phi \circ O_N(\succ_\mathcal{P}) = 1 \text{ and } \phi \circ O_N(\prec_\mathcal{P}) = 0\}$$

Example 2. *(ex. 1 continued) In the following table, we detail all the relevant observed coalitions. Sufficient coalitions appear in the upper right side, boldfaced, while insufficient*

coalitions are in the lower left side. N stands for unanimity, which is self-explanatory.

	$\star\star\star$		$\star\star$		$\star$	
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	–	–	**abc**	**abd**	**abc**	**abd**
A_2	–	–	**N**	**abd**	**N**	**abd**
B_1	d	bd	–	–	**abd**	**abd**
B_2	acd	ac	–	–	**abc**	**abd**
C_1	d	d	acd	bd	–	–
C_2	cd	c	c	bcd	–	–

2.4 Consistency of Judgment

The set $\Phi(\mathcal{P})$ is empty if, and only if, Pareto dominance is contradicted ($\exists\ x^\star, y^\star \in \mathbb{X}^\star,\ \forall i \in N,\ x_i^\star \geq y_i^\star$ and $Class(x^\star) < Class(y^\star)$), or some coalition of criteria $M \in \mathbb{B}^N$ observed as being sufficient is weaker (for inclusion) than some coalition $M' \in \mathbb{B}^N$ observed as being insufficient. In such a case, we call the preference information $\mathcal{P}$ *inconsistent*; otherwise, it is consistent and $\Phi(\mathcal{P})$ is a *partially defined Boolean function* [Crama *et al.*, 1988]. Combining the constraints on the judgment functions expressed by $\widehat{\Phi}$ and by $\mathcal{P}$, we can compute the true points of $\Phi(\mathcal{P})$. They are the antecedents of 1 common to every judgment function $\phi \in \Phi(\mathcal{P})$, and represent the coalitions *established as sufficient*, by the virtue of being at least as strong as an observed sufficient coalition.

$$\mathcal{T}_\mathcal{P} := \{t \in \mathbb{B}^N : \exists\ t_{obs} \in O_N(\succ_\mathcal{P}),\ t_{obs} \subseteq t\}$$

Conversely, the false points are the antecedents of zero common to every $\phi \in \Phi(\mathcal{P})$ and represent the coalitions established as insufficient.

$$\mathcal{F}_\mathcal{P} := \{f \in \mathbb{B}^N : \exists\ f_{obs} \in O_N(\prec_\mathcal{P}),\ f_{obs} \supseteq f\}$$

Proposition 1 details three manners to express inconsistency:

Proposition 1. *For any $\mathcal{P} \subset \mathbb{X} \times \mathbb{K}$, the three following conditions are equivalent and characterize inconsistency:*

1. *Absence of compatible judgment function: $\Phi(\mathcal{P}) = \emptyset$*

2. *Conflicting constraints: $\mathcal{T}_\mathcal{P} \cap \mathcal{F}_\mathcal{P} \neq \emptyset$*

3. *Explicit contradiction: $\exists t \in O_N(\succ_\mathcal{P}), \exists f \in O_N(\prec_\mathcal{P}) : t \subseteq f$*

Example 3. *(ex. 2 continued) Coalitions are sorted according to the observations, and monotonicity:*
$O_N(\succ_\mathcal{P}) = \{N, abc, abd\} = \mathcal{T}_\mathcal{P}$
$O_N(\prec_\mathcal{P}) = \{c, d, ac, bd, cd, acd, bcd\}$
$\mathcal{F}_\mathcal{P} = \{\emptyset, a, b, c, d, ac, ad, bc, bd, acd, bcd\}$
There is no dispute, as $\mathcal{T}_\mathcal{P} \cap \mathcal{F}_\mathcal{P} = \emptyset$, but the coalition ab is left undecided.

3 Recommendations and Explanations

In the previous section, we saw how the decision-maker interprets pairwise comparisons between reference objects belonging to different classes as sufficient or insufficient coalitions of criteria. Here comes a new candidate, $z \in \mathbb{X}$. It gauges every reference object in $\mathbb{X}^\star$, yielding $|\mathcal{P}|$ observations $\overrightarrow{o}(z, \mathcal{P}) := \bigcup_{x^\star \in \mathbb{X}^\star} O_N(z, x^\star)$, and is also evaluated by every reference object, yielding $|\mathcal{P}|$ other observations $\overleftarrow{o}(z, \mathcal{P}) := \bigcup_{x^\star \in \mathbb{X}^\star} O_N(x^\star, z)$. Each of these $2|\mathcal{P}|$ observations is interpreted as a *sufficient, insufficient* or *undecided* coalition of criteria.

Example 4. *(ex. 3 continued) The following table augments the one presented in example 2 with the coalitions resulting from comparisons between the reference objects $A_1, A_2, B_1, B_2, C_1, C_2$ and the candidates X, Y.*

	$\star\star\star$		$\star\star$		$\star$		?	?
	A_1	A_2	B_1	B_2	C_1	C_2	X	Y
A_1	–	–	**abc**	**abd**	**abc**	**abd**	**N**	**N**
A_2	–	–	**N**	**abd**	N	**abd**	**N**	acd
B_1	d	bd	–	–	**abd**	**abd**	**N**	ad
B_2	acd	ac	–	–	**abc**	**abd**	acd	acd
C_1	d	d	acd	bd	–	–	acd	ad
C_2	cd	c	c	bcd	–	–	cd	cd

X	d	b	(ab)	bd	(ab)	**abd**
Y	bd	b	**abc**	bd	**abc**	**abd**

Non-bracketed coalitions have already been sorted according to the preference information: boldfaced coalitions are those previously established as sufficient, the others are insufficient. Bracketed coalitions are yet undecided. $\forall z \in \{X, Y\}$, $\overrightarrow{o}(z, \mathcal{P})$ appears in the corresponding line, and $\overleftarrow{o}(z, \mathcal{P})$ in the appropriate column.

In this section, we specify the mapping between these observations and the output of the classifier system, the *recommendation* $K(z, \mathcal{P}) \subset \mathbb{K}$ and an *explanation* $\mathcal{E}(k, \mathcal{P})$ supporting it.

3.1 Possible Assignments

As defined by the works of [Greco *et al.*, 2010] about *necessary* and *possible* preference relations, the definition of *possible assignments* is closely related to the notion of *consistency* of an assignment with respect to the corpus of preference information. Defining, as we did in Section 2, $\Phi(\mathcal{P})$ as the set of preference parameters compatible to $\mathcal{P}$, and assuming it is not empty:

- *necessary* assignments are yielded by *every* possible completion of these preference parameters;
- *possible* assignments are yielded by *some* possible completion of these preference parameters;
- *impossible* assignments are yielded by *no* possible completion of these preference parameters;

These sets of assignments are concisely described referring to the set:

$$\widehat{K}(z, \mathcal{P}) := \{k \in \mathbb{K} : \Phi(\mathcal{P} \cup \{(z, k)\}) \neq \emptyset\}$$

A possible assignment is in $\widehat{K}(z, \mathcal{P})$, an impossible one is not. When $\widehat{K}(z, \mathcal{P})$ boils down to a singleton, then it is a necessary assignment for z.

This definition of *possible assignment* is straightforward to implement, simply iterating through the set of possible assignments classes $k \in \mathbb{K}$, updating the preference information $\mathcal{P}' \leftarrow \mathcal{P} \cup \{(z, k)\}$, and checking the consistency of $\mathcal{P}'$. Unfortunately, it is a tricky notion when it comes to explaining. The actual unveiling of a Boolean judgment function compatible to the assignment is not very appealing, as it introduces at the same time elements of *jargon* –describing the judgment of the decision-maker as the partition of coalitions of criteria between sufficient and insufficient– and *arbitrariness*, as some coalitions may very well be undecided

and should remain so. Consequently, we adopt the following principle: *"Everything is possible, unless proven otherwise"*.

Doing so shifts the burden of proof towards impossibility, focusing on the exhibition of constraints restricting the set $\widehat{K}(z, \mathcal{P})$. We aim at *explaining* these constraints thanks to *statements* of the form $[premises : conclusions]_{scheme}$. We define several *argument schemes*, as formalized by [Walton, 1996] in order to capture stereotypical patterns of human reasoning. These schemes specify the nature and conditions imposed to both premises and conclusions, yielding to valid arguments. We are looking for *complete* explanations, so we must ensure the validity of the implication $premises \Rightarrow conclusions$, and provide *grounded* sets of statements, such that any premise is either the conclusion of another argument, or directly referencing the assumed available information (pairwise comparisons between the reference objects or the candidate, based on criteria or assignment).

In order to make apparent the cause of impossibility, we consider the potential consequences of assigning a candidate to a class through the *additional (in)sufficient coalitions conditional to the assignment of the candidate z to the class k*:

$$\Delta\mathcal{T}_\mathcal{P}(z, k) := \mathcal{T}_{\mathcal{P} \cup \{(z,k)\}} \setminus \mathcal{T}_\mathcal{P}; \quad \Delta\mathcal{F}_\mathcal{P}(z, k) := \mathcal{F}_{\mathcal{P} \cup \{(z,k)\}} \setminus \mathcal{F}_\mathcal{P}$$

We rewrite the impossibility of assigning the candidate z to the class k using the *conflicting constraints* characterization of inconsistency (see Prop. 1). We consider three potential sources of impossibility, sorted by evidence: $\widehat{K}(z, \mathcal{P}) = \bigcap_{i \in \{1,2,3\}} K_i(z, \mathcal{P})$ where:

- $K_1(z, \mathcal{P}) := \{k \in \mathbb{K} : \mathcal{T}_\mathcal{P} \cap \Delta\mathcal{F}_\mathcal{P}(z, k) = \emptyset\}$ highlights conflicts between established sufficient coalitions, and the assignment of z;
- $K_2(z, \mathcal{P}) := \{k \in \mathbb{K} : \Delta\mathcal{T}_\mathcal{P}(z, k) \cap \mathcal{F}_\mathcal{P} = \emptyset\}$ highlights conflicts between established insufficient coalitions, and the assignment of z;
- $K_3(z, \mathcal{P}) := \{k \in \mathbb{K} : \Delta\mathcal{T}_\mathcal{P}(z, k) \cap \Delta\mathcal{F}_\mathcal{P}(z, k) = \emptyset\}$ takes into account the least obvious situation where some assignment of z may be self-contradictory, without conflicting with any previously acknowledged information.

The next section details the impossibilities captured by the set $K_1(z, \mathcal{P})$, and proposes a supporting explanation $\mathcal{E}_1(z, \mathcal{P})$, while the other cases are briefly presented in section 3.3.

3.2 Assignments Contradicting Previously Established Sufficient Coalitions

In this section, we focus on the set $K_1(z, \mathcal{P}) := \{k \in \mathbb{K} : \mathcal{T}_\mathcal{P} \cap \Delta\mathcal{F}_\mathcal{P}(z, k) = \emptyset\}$. As seen in the previous section this set provides a range of possible assignments for the candidate z, and partially implements the model described by the manifesto exposed in the introduction. We first describe $K_1(z, \mathcal{P})$ as an intersection of constraints, for which we provide a description based on arguments. We prove $K_1(z, \mathcal{P})$ is an interval of $\mathbb{K}$, and provide a short, yet complete, explanation accounting for this recommendation.

For increased readability, we introduce notations for particular sets of classes. For $k \in \mathbb{K}$, let $\mathbb{K}_{\precsim k}$ (resp. $\mathbb{K}_{\succsim k}$) the interval of classes not greater (resp. not lower) than k.

By construction, the recommended set $K_1(z, \mathcal{P})$ is built in order to reject some impossible assignments. To illustrate and understand its behavior, we make up a situation that specifically triggers this rejection flag. Suppose we know that:

(1) the coalition of criteria $T \in \mathbb{B}^N$ is already known to be sufficient, and

(2) the candidate $z \in \mathbb{X}$ is at least as good as the reference object $\underline{x}^\star \in \mathbb{X}^\star$, assigned to class $\underline{k} \in \mathbb{K}$, for all criteria in T.

Then, z outranks $\underline{x}^\star$ and cannot be assigned to a class strictly worse than $\underline{k}$ by application of (R1). This constraint is captured by the set $K_1(z, \mathcal{P})$, as the assignment of z to any class $k \prec \underline{k}$ would lead to conclude that the coalition of criteria $O_N(z, \underline{x}^\star)$ is insufficient, so that the coalition of criteria T would belong to both sets $\Delta \mathcal{F}_\mathcal{P}(z, k)$ and $\mathcal{T}_\mathcal{P}$. Consequently, $k \notin K_1(z, \mathcal{P})$.

If we replace the assumption (2) by:

(2') the reference object $\overline{x}^\star \in \mathbb{X}^\star$, assigned to class $\overline{k} \in \mathbb{K}$, is at least as good as the candidate $z \in \mathbb{X}$ for all criteria in T.

then $\overline{x}^\star \in \mathbb{X}^\star$ outranks z and z cannot be assigned to a class strictly better than $\overline{k}$, as

$$k \succ \overline{k} \Rightarrow \mathcal{T}_\mathcal{P} \ni T \subseteq O_N(\overline{x}^\star, z) \in \Delta \mathcal{F}_\mathcal{P}(z, k) \Rightarrow k \notin K_1(z, \mathcal{P})$$

Reciprocally, any assignment $k_0 \notin K_1(z, \mathcal{P})$ results in a non-empty intersection $\mathcal{T}_\mathcal{P} \cap \Delta \mathcal{F}_\mathcal{P}(z, k_0)$, which involves at least one sufficient coalition $T \in \mathcal{T}_\mathcal{P}$, as in assumption (1), and one stronger, insufficient coalition resulting either from the observations $\overrightarrow{o}(z, \mathcal{P})$, as in assumption (2), or from $\overleftarrow{o}(z, \mathcal{P})$, as in (2').

A statement of type (1) needs to be backed by evidence, so we introduce two argument schemes:

Definition 1. *For any reference objects $a^\star, b^\star \in \mathbb{X}^\star$ and any coalition of criteria $T \in \mathbb{B}^N$, we say the tuple $[a^\star, b^\star : T]_\mathcal{T}$ instantiates the argument scheme* SUFFICIENT COALITION$(\mathcal{P})$ *if, and only if, $T \supseteq O_N(a^\star, b^\star)$ and $a^\star \succ_\mathcal{P} b^\star$. We also say the tuple $[\emptyset : N]_1$ instantiates the argument scheme* WEAK DOMINANCE.

Proposition 2 (Argumentative structure of the sufficient coalitions)**.**

$$\mathcal{T}_\mathcal{P} = \{N\} \cup \bigcup_{[a^\star, b^\star : T]_\mathcal{T}} \{T\}$$

The sufficient coalitions are exactly the conclusions of the arguments instantiating the SUFFICIENT COALITION$(\mathcal{P})$ *scheme.*

In order to account for the atoms of reasoning (2) and (2') and present them to the recipient of the recommendation, we define the corresponding argument schemes.

Definition 2. *For any coalition of criteria $T \in \mathbb{B}^N$, any reference object $x^\star \in \mathbb{X}^\star$ and any class $c \in \mathbb{K}$, we say that:*

- *the tuple $[T, x^\star : \mathbb{K}_{\succsim c}]_{\mathcal{T}/\overrightarrow{o}}$ instantiates the argument scheme* OUTRANKING$(z, \mathcal{P})$ *if, and only if, $T \in \mathcal{T}_\mathcal{P}$ and $\forall i \in T, z_i \geq x_i^\star$ and $class(x^\star) = c$.*

- *the tuple $[T, x^\star : \mathbb{K}_{\precsim c}]_{\mathcal{T}/\overleftarrow{o}}$ instantiates the argument scheme* OUTRANKED$(z, \mathcal{P})$ *if, and only if, $T \in \mathcal{T}_\mathcal{P}$ and $\forall i \in T, x_i^\star \geq z_i$ and $class(x^\star) = c$*

Proposition 3 (Argumentative structure of the recommendation)**.**

$$K_1(z, p) = \mathbb{K} \cap \bigcap_{[T, \underline{x}^\star : \underline{k}]_{\mathcal{T}/\overrightarrow{o}}} \mathbb{K}_{\succsim \underline{k}} \cap \bigcap_{[T, \overline{x}^\star : \overline{k}]_{\mathcal{T}/\overleftarrow{o}}} \mathbb{K}_{\precsim \overline{k}}$$

Proposition 3 is a concise rewording of the necessary and sufficient conditions for a given class *not* to belong to the set $K_1(z, \mathcal{P})$ detailed previously. As a corollary, it shows that $K_1(z, \mathcal{P})$ is an interval of $\mathbb{K}$. Consequently, $K_1(z, \mathcal{P})$ can be completely described by a pair $(\underline{k_1}, \overline{k_1})$ such that:

- $K_1(z, \mathcal{P}) = \mathbb{K}_{\succsim \underline{k_1}} \cap \mathbb{K}_{\precsim \overline{k_1}}$

- the lower bound $\underline{k_1}$ is *maximal*, as there is no class strictly better than $\overline{k_1}$ which is supported by an argument instantiating the OUTRANKING$(z, \mathcal{P})$ scheme. It is *trivial* if $\underline{k_1} = \min \mathbb{K}$ (either when the set OUTRANKING$(z, \overline{\mathcal{P}})$ is empty, or when it does not support a stronger outcome), in which case it does not need any explanation. If $\underline{k_1} \succ \min \mathbb{K}$, then it admits at least one *explanation* $\underline{E_1}$ composed of an argument $[T, \underline{x}^\star : \mathbb{K}_{\succsim \underline{k_1}}]_{\mathcal{T}/\overrightarrow{o}} \in$ OUTRANKING backed by an argument $[a^\star, b^\star : T]_\mathcal{T} \in$ SUFFICIENT COALITION;

- the upper bound $\overline{k_1}$ is *minimal*, as there is no class strictly worse than $\overline{k_1}$ which is supported by an argument instantiating the OUTRANKED$(z, \mathcal{P})$ scheme. It is *trivial* if $\overline{k_1} = \max \mathbb{K}$, in which case it does not need any explanation. If $\overline{k_1} \prec \max \mathbb{K}$, then it admits at least one *explanation* $\overline{E_1}$ composed of an argument $[T', \overline{x}^\star : \mathbb{K}_{\precsim \overline{k_1}}]_{\mathcal{T}/\overleftarrow{o}} \in$ OUTRANKED backed by an argument $[a^\star, b^\star : T']_\mathcal{T} \in$ SUFFICIENT COALITION.

Finally, the recommended interval $K_1(z, \mathcal{P})$ is supported by an explanation $\mathcal{E}_1$ in the form of a pair $(\underline{E_1}, \overline{E_1})$, where $\underline{E_1}$ and $\overline{E_1}$ can be either the empty set or a pair of arguments. Taken together, all these 0, 2 or 4 arguments prove that any assignment $k \in \mathbb{K} \setminus K_1(z, \mathcal{P})$ should be rejected as "impossible". Such explanation is not necessarily unique, and we denote by $\widehat{\mathcal{E}}_1(z, \mathcal{P})$ the set of suitable explanations.

Example 5. *(ex. 4 continued)*
Using the table presented in Example 4, the set K_1 can be interpreted as "a candidate cannot be assigned a class laying strictly on the right of, nor a class strictly above, a case containing a boldfaced coalition": Consequently,

- $\begin{cases} K_1(X, \mathcal{P}) = \{\star, \star\star\} \\ \mathcal{E}_1(X, \mathcal{P}) \ni (\emptyset, \{[\emptyset : N]_1, [N, B_1 : \precsim \star\star]_{\mathcal{T}/\overleftarrow{o}}\}) \end{cases}$
  X cannot be ranked higher than $\star\star$, because B_1 is rated $\star\star$ and dominates X.

- $\begin{cases} K_1(Y, \mathcal{P}) = \{\star\star, \star\star\star\} \\ \widehat{\mathcal{E}}_1(Y, \mathcal{P}) \ni (\{[A_1, C_1 : abc]_\mathcal{T}, [abc, B_1 : \succsim \star\star]_{\mathcal{T}/\overrightarrow{o}}\}, \emptyset) \end{cases}$
  Y cannot be ranked lower than $\star\star$, because it outranks B_1. Indeed, Y compares to B_1 the same way as A_1 to C_1: it is at least as good on the sufficient coalition of criteria abc.

3.3 Other Impossible Assignments

The set $K_2(z, \mathcal{P})$ is defined symmetrically from $K_1(z, \mathcal{P})$ w.r.t. sufficient and insufficient coalitions. Assignments *not* in $K_2(z, \mathcal{P})$ result from the collision of a coalition of criteria known to be insufficient, and the observation of a candidate object resulting in an even weaker coalition, so outranking is excluded, and all the classes strictly above or below (depending on the direction of observation) the one of the reference object are therefore forbidden. *Mutatis mutandis*, we can define the argument schemes INSUFFICIENT COALITION($\mathcal{P}$), WEAKLY DOMINATED, NOT OUTRANKING($z, \mathcal{P}$), NOT OUTRANKED($z, \mathcal{P}$) and obtain the same structural results, leading to define similar explanations for the lower and upper bounds of the interval $K_2(z, \mathcal{P})$.

Example 6. *(ex. 4 continued)*

Using the table presented in Ex. 4, the set K_2 interprets the insufficient coalitions of the table, those not boldfaced nor parenthesized. A candidate cannot be assigned a class strictly below, nor strictly on the left, of such cases. For instance, $O_N(B_2, X) = acd \in \mathcal{F}_\mathcal{P}$ (e.g. because $O_N(C_1 \prec_\mathcal{P} B_1) = acd$), so X is not outranked by B_2 and should be at least assigned the same class ($\star\star$), and $O_N(X, B_2) = bd \in \mathcal{F}_\mathcal{P}$ (e.g. because it is weaker than $bcd = O_N(C_2 \prec_\mathcal{P} B_2)$), so X does not outrank B_2 and should not be assigned a strictly better class ($\star\star$). In terms of preference, objects X and B_2 are incomparable, and thus should be assigned the same class. Finally, $K_2(X, \mathcal{P}) = \{\star\star\}$.

The set $K_3(z, \mathcal{P})$ excludes inconsistent judgments on yet undecided coalitions of criteria. There is no guarantee that $K_3(z, \mathcal{P})$ has an interval structure. We omit this case due to space limitations.

4 An argumentative Perspective

Along this paper, we proposed the construction of explanations supporting results of a multi-criteria sorting problem, as combinations of arguments schemes. Each instantiation of one of the six previous main schemes (see Def. 1, 2 and their symmetrical forms) provides one type of argument. These arguments may be conflicting, and two different relations can be distinguished:

Conflicting coalitions: we have evidence indicating that a given coalition is potentially at the same time sufficient and insufficient (i.e. there are two coalitions $t \subseteq f$ such that $[a^\star, b^\star : t]_\mathcal{T}$ and $[c^\star, d^\star : f]_\mathcal{F}$). This situation represents an explicit contradiction corresponding to an inconsistency situation (see Sec. 2.4). Such conflicts are not illustrated through the previous examples, however inconsistencies are classical situations within decision problems, as it concerns a human decision-maker.

Conflicting classification: it may occur that, for some candidate, arguments based on the outranking relation point towards an *empty* interval of possible assignments. This situation corresponds to the fact that the sets $K_1(z, \mathcal{P})$ and $K_2(z, \mathcal{P})$ are disjoint, which may happen when either is empty, or when the lower bound of one exceed the upper bound of the other.

Example 7. *(ex. 4 cont.) Y and A_2 are incomparable, Y and B_2 are incomparable, yet A_2 is preferred to B_2. In particular, A_2 ($\star\star\star$) does not outrank Y and Y does not outrank B_2 ($\star\star$) so $K_2(Y, \mathcal{P}) = \emptyset$.*

The impossibility to provide any recommendation is clearly critical from the point of view of decision aiding. These unfortunate situations cannot be ruled out in the general case, as they may stem from Condorcet paradoxes (failures of transitivity) concerning the necessary outranking relation or the necessary not-outranking relation (see e.g. [Köksalan *et al.*, 2009] for a discussion).

The argumentative treatment for our multi-criteria ordinal sorting problem is thus to construct arguments pro and against each possible assignment (of the reference object and the candidate), and to determine among conflicting arguments the *acceptable* ones. This can be done by taking two different perspectives. One way is to rely on the work of [Dung, 1995] - the next question being to identify which semantics are appropriate in our situation. This is close in spirit to an approach presented in [Amgoud and Serrurier, 2007] for classification in *unordered* classes (however in our context the relation between arguments would be symmetric [Coste-Marquis *et al.*, 2005]). Another perspective is to consider the construction of the argumentation system as a dialogue game and to rely on critical questions [Walton, 1996; Ouerdane *et al.*, 2008] to evaluate the arguments. This perspective has the advantage to keep the decision-maker in the loop, which is often essential in a decision situation [Labreuche *et al.*, 2015]. Both approaches look promising and are made possible thanks to the modeling presented in this paper.

5 Conclusion

We have presented a fully accountable multi-criteria ordinal sorting model, based on several design principles and assumptions. The strength of the model is that it solely relies on a simple set of classification rules, which means that each recommendation can be justified by instantiating and combining these rules–nothing else. Several argument schemes have been proposed for that purpose. Interestingly, some of these schemes have a flavour of analogical reasoning, which was studied in the context of classification [Hug *et al.*, 2016]. Now the simplicity of our model comes at a price: there are different situations where inconsistency might occur, and the model is not equipped yet to handle such situations. Facing this issue we can take two stances. The first one is to relax some of our design assumptions. For instance, we may decide that it is actually acceptable for the model to use a *frontier* between classes (allowing to eschew the Condorcet paradox). This would require original explanation techniques to maintain the desired accountability. Another avenue is to handle the inconsistencies thanks to defeasible and non-monotonic reasoning techniques [Brewka *et al.*, 2008]. Our discussion in Section 4 points to formal argumentation as a natural and promising opportunity for future research.

References

[Amgoud and Serrurier, 2007] Leila Amgoud and Mathieu Serrurier. Arguing and explaining classifications. In *Proceeding of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 160, 2007.

[Belahcene *et al.*, 2017] Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ourdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183, 2017.

[Bouyssou and Marchant, 2007] Denis Bouyssou and Thierry Marchant. An axiomatic approach to noncompensatory sorting methods in mcdm, i: The case of two categories. *EJOR*, 178(1):217–245, 2007.

[Bouyssou, 1986] Denis Bouyssou. Some remarks on the notion of compensation in mcdm. *EJOR*, 26(1):150–160, 1986.

[Brewka *et al.*, 2008] Gerhard Brewka, Ilkka Niemelä, and Miroslaw Truszczynski. Nonmonotonic reasoning. In *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 239–284. Elsevier, 2008.

[Burell, 2016] Jenna Burell. How the machine "thinks": understanding opacity in machine learning algorithms. *Big Data and Society*, 1(3), 2016.

[Coste-Marquis *et al.*, 2005] Sylvie Coste-Marquis, Caroline Devred, and Pierre Marquis. Symmetric argumentation frameworks. In *Proceedings of the 8th European Conference Symbolic and Quantitative Approaches to Reasoning with Uncertainty ECSQARU*, pages 317–328. Springer, 2005.

[Crama *et al.*, 1988] Yves Crama, Peter L. Hammer, and Toshihide Ibaraki. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research*, 16(1):299–325, 1988.

[Datta *et al.*, 2016] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *The 37th IEEE Symposium on Security and Privacy (Oakland)*, 2016.

[Dias *et al.*, 2002] Luis Dias, Vincent Mousseau, José Figueira, and Joao Clímaco. An aggregation / disaggregation approach to obtain robust conclusions with electre tri. *EJOR*, 138(2):332–348, 2002.

[Dung, 1995] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[Fishburn, 1976] Peter C. Fishburn. Noncompensatory preferences. *Synthese*, 33(2/4):393–403, 1976.

[Goodman and Flaxman, 2016] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". ArXiv e-prints: 1606.08813, June 2016.

[Greco *et al.*, 2008] Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *EJOR*, 191(2):416–436, 2008.

[Greco *et al.*, 2010] Salvatore Greco, Roman Słowinski, José Figueira, and Vincent Mousseau. Robust ordinal regression. In *Trends in Multiple Criteria Decision Analysis*, pages 241–284. Springer Verlag, 2010.

[Hug *et al.*, 2016] Nicolas Hug, Henri Prade, Gilles Richard, and Mathieu Serrurier. Analogical classifiers: A theoretical perspective. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence*, pages 689–697, 2016.

[Keeney and Raiffa, 1976] Ralph L. Keeney and Howard Raiffa. *Decisions with multiple objectives: Preferences and value tradeoffs*. J. Wiley, New York, 1976.

[Köksalan *et al.*, 2009] Murat Köksalan, Vincent Mousseau, Ozgur Ozpeynirci, and Selin Bilgin Ozpeynirci. A new outranking-based approach for assigning alternatives to ordered classes. *Naval Research Logistics*, 56(1):74–85, 2009.

[Labreuche *et al.*, 2012] Christophe Labreuche, Nicolas Maudet, and Wassila Ourdane. Justifying Dominating Options when Preferential Information is Incomplete. In *ECAI'12*. IOS Press, 2012.

[Labreuche *et al.*, 2015] Christophe Labreuche, Nicolas Maudet, Wassila Ourdane, and Simon Parsons. A dialogue game for recommendation with adaptive preference models. In *Proceedings of the 14th International Conference on Autonomous Agent and MultiAgent systems (AAMAS)*, pages 959–967, 2015.

[Leroy *et al.*, 2011] Agnes Leroy, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a multiple criteria sorting method. In *ADT*, pages 219–233. Springer, 2011.

[Ourdane *et al.*, 2008] Wassila Ourdane, Nicolas Maudet, and Alexis Tsoukiàs. Argument schemes and critical questions for decision aiding process. In *COMMA*, pages 285–296. IOS Press, 2008.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[Roy, 1991] Bernard Roy. The outranking approach and the foundations of electre methods. *Theory and decision*, 31(1):49–73, 1991.

[Tintarev, 2007] Nina Tintarev. Explanations of recommendations. In *Proc. ACM conference on Recommender systems*, pages 203–206, 2007.

[Vincke, 1999] Philippe Vincke. Robust solutions and methods in decision-aid. *Journal of multicriteria decision analysis*, 8(3):181, 1999.

[Walton, 1996] Douglas Walton. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996.

Accountable Approval Sorting

Khaled Belahcene[1], Yann Chevaleyre[2], Christophe Labreuche[3],
Nicolas Maudet[4], Vincent Mousseau[1], Wassila Ouerdane[1]

[1] Laboratoire Genie Industriel, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
[2] Université Paris-Dauphine, PSL Research University, CNRS, UMR [7243], LAMSADE, France
[3] Thales Research and Technology, Palaiseau, France
[4] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, France
{khaled.belahcene, vincent.mousseau, wassila.ouerdane}@centralesupelec.fr,
yann.chevaleyre@dauphine.fr, christophe.labreuche@thalesgroup.com, nicolas.maudet@lip6.fr

Abstract

We consider decision situations in which a set of points of view (voters, criteria) are to sort a set of candidates to ordered classes (GOOD / BAD). Candidates are judged GOOD when approved by a sufficient set of points of view; this corresponds to noncompensatory sorting. To be accountable, such approval sorting should provide guarantees about the decision process and decisions concerning specific candidates. We formalize accountability using a feasibility problem expressed as a boolean satisfiability formulation. We illustrate different forms of accountability when a committee decides with approval sorting and study the information that should be disclosed by the committee.

1 Introduction

A committee meets to decide upon the sorting of a number of candidates into two categories (e.g. candidates to accept or not, projects to fund or not). The committee applies a decision process which is public, the outcomes are public as well, however the details of the votes are sensitive and should not be made available. Recently, the issue of the *accountability* of algorithmic decisions has become a primary concern of our society [Doshi-Velez *et al.*, 2017; Wachter *et al.*, 2017]. To what extent can we make the committee accountable of its decisions? In particular, in our setting, a distinctive feature is that the decision may concern several individuals: being accountable for the classification of an individual may not be the same as being accountable for all the classifications. To make things more precise, it is thus useful to distinguish the following situations:

S1: an independent audit agency is commissioned to check that the decisions of the committee indeed comply with the publicly announced decision rule.

S2: a candidate, (supposedly) unsatisfied with the outcome of the process regarding his own classification, challenges the committee and asks for a justification.

Situation S1 is sometimes called *procedural regularity*, see for instance [Kroll *et al.*, 2017], which calls for systems able to prove to oversight authorities that "decisions are made under an announced set of rules consistently applied in each case". A typical way to address situation S1 is to require *transparency* and let the audit agency access all the available information. This suffers from two drawbacks: (i) there are often exceptions making full disclosure of the decision procedure impossible, (ii) the burden of proof lies on the shoulders of the audit agency, which (depending on the model) may be too demanding. Alternatively, we can leave the burden of proof on the committee's side and ask for evidence that the set of classifications is compliant with the decision process. This may be done by exhibiting only part of the information, illustrating that the obtained classification is a *possible* outcome of the sorting process. Since, typically, many other outcomes would also be possible, this could preserve to some extent the privacy of the committee's votes. On the other hand, failing this test would be evidence that the process was biased.

Regarding situation S2, the objective is to justify the classification of the complaining individual, again with minimal disclosure of the committee's votes. In this case, the committee will aim for evidence that the classification of the candidate cannot be otherwise, *as long as a number of other classification outcomes are accepted*. We can think of such decisions as reference cases. Technically, this requires to show the impossibility to rank the candidate in a different category, *i.e.* the decision is *necessary* with respect to the jurisprudence.

More precisely, we shall primarily be concerned with a general sorting model where voters express binary judgments [Laslier and Sanver, 2010], and candidates are sorted as either *good* or *bad* depending on the fact that the coalition of voters supporting this classification is winning or not. An important hypothesis is that the set of winning coalitions has to remain constant for the set of classifications under scrutiny. This can be seen as a requirement for the process to be unbiased. In this setting, the "details of the votes" cover two aspects: (i) the approval of voters at the individual level, (ii) the winning coalitions at the committee level. In this paper we address the following research question:

Can we make the decisions of a committee using approval sorting accountable while preserving as much as possible the details of the votes?

The details of the sorting model are given in Section 2. At the core of our proposal lies a characterization result of the sorting model which avoids explicit reference to winning coalitions, and leads to a SAT encoding (Section 3). In Section 4, we consider the different scenarios discussed in the introduction and show how this formal machinery allows us to provide argument schemes which answer, at least partially, the accountability requirements. Section 5 discusses related work and concludes.

2 Noncompensatory Sorting

We are interested in situations where there is a need to aggregate diverse, potentially conflicting, *points of view* forming a set $\mathcal{N}$ – each $i \in \mathcal{N}$ can be seen as an agent, a voter, or a criterion – into a single *sorting* of some *alternatives* taken in a set $\mathbb{X}$ between two categories, GOOD and BAD, expressed by an *assignment* $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$. Each point of view $i \in \mathcal{N}$ has an opinion on the entire set of alternatives in the form of a complete preorder $\succsim_i$ (*i.e.* $\succsim_i$ is a complete, reflexive and transitive binary relation on $\mathbb{X}$). This preference may stem from numeric or symbolic performance, as it is often the case in multi-criteria decision aiding, or be intrinsically ordinal, as it is often assumed in social choice contexts. Nevertheless, the aggregation procedure requires that each point of view $i \in \mathcal{N}$ expresses only a binary judgment on each alternative $x \in \mathbb{X}$ which is either approved or not according to i. We shall also consider a subset $\mathbb{X}^\star \subseteq \mathbb{X}$ of alternatives with a *reference* status, with their assignment $\alpha^\star : \mathbb{X}^\star \to \{\text{GOOD}, \text{BAD}\}$ serving as a basis for elaborating justifications.

This abstract description covers several well-documented decision processes, e.g. :

- a multiple criteria sorting problem [Bouyssou *et al.*, 2006] with ordinal preferences (each point of view $i \in \mathcal{N}$ is a *criterion*);
- a committee decision context (each point of view $i \in \mathcal{N}$ is a *voter* and the GOOD category is the set of winners).

Example 1. *We consider a situation with six alternatives* $\mathbb{X} := \{a, b, c, d, e, f\}$, *assessed from five points of view* $\mathcal{N} := \{1, 2, 3, 4, 5\}$ *in the following manner:*

$$a \succ_1 b \succ_1 f \succ_1 e \succ_1 c \succ_1 d$$
$$e \succ_2 b \succ_2 c \succ_2 d \succ_2 a \succ_2 f$$
$$f \succ_3 a \succ_3 b \succ_3 d \succ_3 e \succ_3 c$$
$$d \succ_4 a \succ_4 c \succ_4 e \succ_4 f \succ_4 b$$
$$c \succ_5 e \succ_5 b \succ_5 f \succ_5 d \succ_5 a$$

We recall the definitions of an upset and the upper closure of a subset w.r.t. a binary relation:

Definition 1 (Upset and upper closure). *Let A be a set and $\mathcal{R}$ a binary relation on A. An upset of $(A, \mathcal{R})$ is a subset $B \subseteq A$ such that $\forall a \in A, \forall b \in B, a\mathcal{R}b \Rightarrow a \in B$. The upper closure of a subset of $(A, \mathcal{R})$ is the smallest upset of $(A, \mathcal{R})$ containing it: $\forall B \subseteq A, cl_A^{\mathcal{R}}(B) := \{a \in A : \exists b \in B \ a\mathcal{R}b\}$.*

We postulate that the process is bounded by two assumptions of rationality, individual and collective.

- At the individual level, for all points of view $i \in \mathcal{N}$, the approved subset of alternatives $\mathcal{A}_i \subseteq \mathbb{X}$ should be an upset for the preference relation $\succsim_i$. Hence, there is no pair of alternatives $x, x' \in \mathbb{X}$ where x is preferred to x' w.r.t. $\succsim_i$, x' is approved by i but not x.

- At the collective level, an alternative $x \in \mathbb{X}$ is collectively approved and sorted into the upper category if, and only if, it is approved by a sufficient coalition of points of view. We assume the set of sufficient coalitions $\mathcal{S} \subseteq \mathcal{P}(\mathcal{N})$ is fixed, and is an upset for inclusion. Hence, if a coalition is sufficient, any superset of this coalition is also sufficient (and if a coalition is insufficient, any subset of it is also insufficient). We do not assume the set of sufficient coalitions has an additive structure, as opposed to weighted voting games or approval balloting [Laslier and Sanver, 2010].

These two stages form the noncompensatory sorting model:

Definition 2 (NCS - noncompensatory sorting model, [Bouyssou and Marchant, 2007]). *Given a set of alternatives $\mathbb{X}$, a set of points of view $\mathcal{N}$, and a tuple of complete preorders $\succsim_i, i \in \mathcal{N}$, if $\mathcal{S}$ is an upset of $(\mathcal{P}(\mathcal{N}), \subseteq)$ and a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$, [1] the noncompensatory sorting model with parameters $(\mathcal{S}, \langle \mathcal{A}_i \rangle)$ is the function $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ mapping alternatives from $\mathbb{X}$ to categories in $\{\text{GOOD}, \text{BAD}\}$ such that the alternative x is assigned to the upper category GOOD if, and only if, the set of points of view according to which x is approved is sufficient, i.e.*

$$NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(x) = \begin{cases} \text{GOOD}, & \text{if } \{i \in \mathcal{N} : x \in \mathcal{A}_i\} \in \mathcal{S} \\ \text{BAD}, & \text{else} \end{cases}$$

$\mathcal{S}$ is the set of sufficient coalitions of the model, and each $\mathcal{A}_i$ is the approved set according to the point of view $i \in \mathcal{N}$.

Example 2. *(ex. 1 continued) Suppose the approved sets are as follows: $\mathcal{A}_1 := \{a, b, f\}, \mathcal{A}_2 := \{e, b, c\}, \mathcal{A}_3 := \{f, a, b\}, \mathcal{A}_4 := \{d, a, c\}, \mathcal{A}_5 := \{c, e, b\}$, corresponding to the three best alternatives according to the respective points of view (3-approval). Suppose also the points of view are aggregated according to the simple majority rule, i.e. $B \in \mathcal{S} \iff |B| \geq 3$. Then, the corresponding noncompensatory model assigns a, b, c to the GOOD category, and d, e, f to the BAD one. Hence, $\alpha := \{(a, \text{GOOD}), (b, \text{GOOD}), (c, \text{GOOD}), (d, \text{BAD}), (e, \text{BAD}), (f, \text{BAD})\}$. We note the same assignment α can be obtained with different sorting parameters, e.g. approved sets $\mathcal{A}'_1 := \{a, b, f\}, \mathcal{A}'_2 := \{e, b, c, d, a\}, \mathcal{A}'_3 := \{\}, \mathcal{A}'_4 := \{d, a, c\}, \mathcal{A}'_5 := \{c\}$ and sufficient coalitions $\mathcal{S}'$ containing the coalitions $\{1, 2\}, \{5\}$ and their supersets.*

This model may appear particularly unwieldy to use explicitly, as it requires to handle a set of sufficient coalitions that lies in the power set of the points of view.

---

[1] Meaning $\langle \mathcal{A}_i \rangle_{i \in \mathcal{N}}$ is a tuple of subsets of $\mathbb{X}$ such that, for all $i \in \mathcal{N}, \mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$. Also, throughout the paper, when the indexing is left unspecified, the tuples are indexed by points of view $i \in \mathcal{N}$.

We propose an indirect approach w.r.t. the parameters of the noncompensatory sorting model implicitly describing the decision process: we suppose the inputs (ordinal preferences over the alternatives according to each point of view) and outputs (an assignment of each alternative to a category, either GOOD or BAD) of the aggregation model are given, and we query the parameters (sufficient coalitions of points of view and accepted sets according to each point of view) of the model. Unlike the usual learning approach, based on the inverse problem of finding the *value* of a suitable tuple of parameters permitting to restore the output given the input, we instead focus on versions of this problem where the issue is merely the *existence* of such a tuple of parameters, and, in the case of a positive answer, to find suitable values for the accepted sets (but not for the set of sufficient coalitions).

Definition 3 (Inverse noncompensatory sorting problem: In-v-NCS). *Given an assignment $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ of alternatives to categories, we say that α can be represented in the noncompensatory sorting model if, and only if, there is a pair of parameters $(\mathcal{S}, \langle \mathcal{A}_i \rangle)$ where $\mathcal{S}$ is an upset of $(\mathcal{P}(\mathcal{N}), \subseteq)$ and $\langle \mathcal{A}_i \rangle_{i \in \mathcal{N}}$ is a tuple of subsets of $\mathbb{X}$ such that, for all $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$, so that $\alpha \equiv NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$.*

We say that α is a *possible assignment* if it is a YES instance of Inv-NCS, *i.e.* α can be represented in the noncompensatory sorting model. When there is some jurisprudence $\alpha^\star$, the assignment of a new candidate x can be *necessary*, in the sense that no other assignment is possible.

Definition 4 (Necessary assignment w.r.t. reference cases). *Given a YES instance $\alpha^\star$ of Inv-NCS, an alternative $x \in \mathbb{X}$ is necessarily assigned to a category $C \in \{\text{GOOD}, \text{BAD}\}$ w.r.t. assignment $\alpha^\star$ if $\alpha^\star \cup \{(x, \overline{C})\}$ is a NO instance of Inv-NCS, where $\overline{C}$ denotes the category opposite to C.*

3 Feasibility of the Inverse NCS Problem

In this section, we propose a characterization of the possibility, given ordinal preferences over the alternatives according to each point of view and an assignment of each alternative to a category, either GOOD or BAD, of representing this assignment in the non-compensatory sorting model. This formulation circumvents any reference to the power set of points of view, so we derive a compact SAT formulation for the inverse problem, which is shown to be NP-hard.

3.1 Inv-NCS with Fixed Approved Sets

When the approved sets are given, solving the inverse NCS problem – *i.e.* learning a set of sufficient coalitions permitting to represent the assignment in the noncompensatory sorting model – is similar to learning a disjunctive normal form from training examples. From this observation, we derive a tractable (computable in polynomial time) algorithm yielding the *version space* [Mitchell, 1982] of the noncompensatory sorting model with fixed approved sets:

Definition 5 (Observed sufficient and insufficient coalitions given approved sets). *Given $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ and a*

tuple $\langle \mathcal{A}_i \rangle$ *of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$, we note:*

$$\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^{\supseteq}_{\mathcal{P}(\mathcal{N})} \left(\bigcup_{g \in \alpha^{-1}(\text{GOOD})} \{i \in \mathcal{N} : g \in \mathcal{A}_i\} \right),$$

$$\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^{\subseteq}_{\mathcal{P}(\mathcal{N})} \left(\bigcup_{b \in \alpha^{-1}(\text{BAD})} \{i \in \mathcal{N} : b \in \mathcal{A}_i\} \right)$$

Proposition 1 (Lower and upper bounds for the sufficient coalitions given the approved sets). *Given an assignment α, a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$ and an upset $\mathcal{S}$ of $(\mathcal{P}(\mathcal{N}), \subseteq)$, α is represented by the noncompensatory sorting model $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ if, and only if:*

$$\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \subseteq \mathcal{S} \subseteq \mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$$

Proof. α is represented by $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ iff i) for all alternatives $g \in \alpha^{-1}(\text{GOOD})$, $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(g) = \text{GOOD}$; and ii) for all alternatives $b \in \alpha^{-1}(\text{BAD})$, $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(b) = \text{BAD}$

i) holds iff $\mathcal{S}$ contains $\bigcup_{g \in \alpha^{-1}(\text{GOOD})} \{i \in \mathcal{N} : g \in \mathcal{A}_i\}$ and, as a consequence of being an upset for inclusion, $\mathcal{S}$ contains $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$. ii) holds iff $\mathcal{S}$ does not contain any coalition pertaining neither to $\bigcup_{b \in \alpha^{-1}(\text{BAD})} \{i \in \mathcal{N} : b \in \mathcal{A}_i\}$ nor to $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$. $\square$

Corollary 1 (complexity of Inv-NCS with fixed approved sets). *Given an assignment α of alternatives to categories and a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$, the problem of deciding whether α can be represented in the noncompensatory sorting model with approved sets $\langle \mathcal{A}_i \rangle$ is tractable (computable in polynomial time).*

Indeed, it boils down to checking whether $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is empty or not, which is $O(|\mathbb{X}|^2 \cdot |\mathcal{N}|)$.

3.2 A Pairwise Formulation for Inv-NCS

The following Theorem is very important as it says that, in order to check that an assignment α is compatible with NCS, it is equivalent to find approval subsets over each point of view such that one can discriminate each pair of GOOD and BAD alternatives on at least one point of view (*i.e.* the GOOD alternative is approved on this point of view, and not the BAD one). Interestingly, the concept of sufficient coalitions disappears in the characterization.

Theorem 1 (Pairwise formulation of the noncompensatory sorting model). *An assignment α of alternatives to categories can be represented in the noncompensatory sorting model if, and only if, there is a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{\mathcal{N}}$ such that:*

1. *for each point of view $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$*

2. *for each pair of alternatives $(g, b) \in \alpha^{-1}(\text{GOOD}) \times \alpha^{-1}(\text{BAD})$, there is at least one point of view $i \in \mathcal{N}$ such that $g \in \mathcal{A}_i$ and $b \notin \mathcal{A}_i$.*

Proof. $[\neg(1+2) \Rightarrow \neg\text{NCS}]$ If there are two alternatives $g \in \alpha^{-1}(\text{GOOD})$ and $b \in \alpha^{-1}(\text{BAD})$ that falsify Condition 2, then, for any potential parameters $\mathcal{S}, \langle \mathcal{A}_i \rangle$ of a noncompensatory sorting model, the nesting $\{i \in \mathcal{N} : g \in \mathcal{A}_i\} \subseteq \{i \in \mathcal{N} : b \in \mathcal{A}_i\}$ results in a sorting $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ at least as favorable to b as to g, whereas $\alpha(b) = \text{BAD}$ is strictly worse than $\alpha(g) = \text{GOOD}$.

[(1+2) $\Rightarrow$ NCS] Given a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{\mathcal{N}}$ satisfying conditions 1 and 2, we consider the sets of coalitions $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$.

According to Proposition 1, α can be represented in the noncompensatory model iff $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) = \emptyset$. Suppose this intersection is nonempty, and let $B \in \mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$. By definition of $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$, there is an alternative $g \in \alpha^{-1}(\text{GOOD})$ such that $B \supseteq \{i \in \mathcal{N} : g \in \mathcal{A}_i\}$: for all points of view $i \notin B$, $g \notin \mathcal{A}_i$. By definition of $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$, there is an alternative $b \in \alpha^{-1}(\text{BAD})$ such that $B \subseteq \{i \in \mathcal{N} : b \in \mathcal{A}_i\}$: for all points of view $i \in B$, $b \in \mathcal{A}_i$. Consequently, there is no point of view according to which g is accepted but not b, contradicting condition 2. Hence, $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) = \emptyset$. $\qquad\square$

3.3 Complexity of Inv-NCS

We show that the inverse NCS problem is intractable.

Proposition 2 (NP-hardness of Inv-NCS).

Given an assignment α of alternatives to categories, the problem of deciding whether α can be represented in the noncompensatory sorting model is NP-hard.

Proof. By reduction from SAT: consider a SAT instance in conjunctive normal form, with n variables $y^1, \ldots, y^n$ and m clauses $c_1 \wedge \cdots \wedge c_m$. We build a gadget assignment with $m + n$ points of view and $2m$ alternatives: $g_1, \ldots, g_m$ are assigned to GOOD whereas $b_1, \ldots, b_m$ are assigned to BAD. First, let us focus on the first m points of view: for each $k \in 1 \ldots m$, let $g_k \sim_k b_k \succ_k g_1 \sim_k \cdots \sim_k g_{k-1} \sim_k g_{k+1} \sim_k \cdots \sim_k g_m \sim_k b_1 \sim_k \cdots \sim_k b_{k-1} \sim_k b_{k+1} \sim_k \cdots \sim_k b_m$. The preference $\succsim_k$ has two equivalence classes, the upper one containing $\{g_k, b_k\}$ and the lower one containing $\bigcup_{k' \neq k} \{g_{k'}, b_{k'}\}$. The n last points of view of the gadget are built considering the SAT formula. From the j-th clause, written in disjunctive form $c_j := \bigvee_{k \in P_j} y^k \vee \bigvee_{k \in N_j} \neg y^k$, where P_j and N_j are disjoint subsets of $1 \ldots n$ indexing the positive (resp. negative) atoms of c_j, we build the preference relation $\succsim_{j+m}$. It has at most 3 equivalence classes: the uppermost containing the alternatives $\bigcup_{k \in P_j} \{g_k\}$, the one in the middle containing $\bigcup_{k \in P_j} \{b_k\} \cup \bigcup_{k \in N_j} \{g_k\}$, and the lowest containing $\bigcup_{k \in N_j} \{b_k\} \cup \bigcup_{k \notin P_j \cup N_j} \{g_k, b_k\}$. We note trivial accepted sets – *i.e.* points of view $i \in \mathcal{N}$ such that $\mathcal{A}_i = \emptyset$ or $\mathcal{A}_i = \mathbb{X}$ – do not contribute to the feasibility of the inverse NCS problem. For the m first points of view, there is only one nontrivial accepted set: it accepts the upper class and rejects the lower one. For the n last points of view of the gadget, the nontrivial accepted sets accept the uppermost equivalence class, reject the lowest class, and either accept or reject the class in the middle. We define a one-to-one mapping between the nontrivial accepted sets of the gadget and the assignment of the n variables of the SAT problem: y^j is False $\iff \bigcup_{k \in P_j} \{b_k\} \cup \bigcup_{k \in N_j} \{g_k\} \in \mathcal{A}_{m+j}$. Each non trivial assignment discriminates all pairs $(g_k, b_{k'})$ with $k \neq k'$ w.r.t. the point of view k. The pairs (g_k, b_k) is discriminated iff the clause c_k is satisfied. Thus, a solution of the SAT problem is mapped to a tuple of accepted sets that discriminates all pairs with opposite assignments and reciprocally. $\qquad\square$

3.4 A Compact SAT Formulation for Inv-NCS

We leverage Theorem 1 by formulating a boolean satisfiability problem that answers the decision problem: can the assignment α be represented in the non-compensatory model? If the instance is a YES, any solution of the satisfiability problem translates into suitable, yet arbitrary, explicit values for the approved sets. Upper and lower bounds for the set of sufficient coalitions can be obtained thanks to Proposition 1.

Corollary 2 (CNF Pairwise SAT formulation for NCS). *Let $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ an assignment. We define the boolean function $\phi_\alpha^{pairwise}$ with variables:*

- *$\lambda_{i,x}$ indexed by a point of view $i \in \mathcal{N}$, and a value $x \in \mathbb{X}$,*
- *$\mu_{i,g,b}$ indexed by a point of view $i \in \mathcal{N}$, a good alternative $g \in \alpha^{-1}(\text{GOOD})$ and a bad alternative $b \in \alpha^{-1}(\text{BAD})$,*

  as the conjunction of clauses: $\phi_\alpha^{pairwise} := \phi_\alpha^1 \wedge \phi_\alpha^2 \wedge \phi_\alpha^3 \wedge \phi_\alpha^4$

$$\phi_\alpha^1 := \bigwedge_{i \in \mathcal{N}} \bigwedge_{x' \succsim_i x} (\lambda_{i,x'} \vee \neg\lambda_{i,x})$$

$$\phi_\alpha^2 := \bigwedge_{i \in \mathcal{N}, \, g \in \alpha^{-1}(\text{GOOD}), \, b \in \alpha^{-1}(\text{BAD})} (\neg\mu_{i,g,b} \vee \neg\lambda_{i,b})$$

$$\phi_\alpha^3 := \bigwedge_{i \in \mathcal{N}, \, g \in \alpha^{-1}(\text{GOOD}), \, b \in \alpha^{-1}(\text{BAD})} (\neg\mu_{i,g,b} \vee \lambda_{i,g})$$

$$\phi_\alpha^4 := \bigwedge_{g \in \alpha^{-1}(\text{GOOD}), \, b \in \alpha^{-1}(\text{BAD})} \left(\bigvee_{i \in \mathcal{N}} \mu_{i,g,b}\right)$$

α can be represented in the noncompensatory sorting model if, and only if, $\phi_\alpha^{pairwise}$ is satisfiable.

Moreover, if $\langle \lambda_{i,x} \rangle, \langle \mu_{i,g,b} \rangle$ is an antecedent of 1 by $\phi_\alpha^{pairwise}$, then the noncompensatory sorting model $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ with accepted sets defined by $\mathcal{A}_i := \{x \in \mathbb{X} : \lambda_{i,x} = 1\}$ and any upset $\mathcal{S}$ of $(\mathcal{P}(\mathcal{N}), \subseteq)$ of sufficient coalitions containing the upset $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and disjoint from the lower set $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ satisfies $\alpha \equiv NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$.

Variables $\lambda_{i,x}$ are assigned to 1 when the alternative x is accepted from the point of view i, and variables $\mu_{i,g,b}$ are assigned to 1 when the point of view i accepts g but not b.

The clauses ϕ_α^1 ensure the sets of accepted values of each point of view meet the first condition of Theorem 1, i.e. $\mathcal{A}_i$ is an upset. The clauses ϕ_α^2 (resp. ϕ_α^3) ensure each variable $\mu_{i,g,b}$ cannot take a value of one unless g is accepted (resp. unless b is not accepted). The clauses ϕ_α^4 ensure the second condition of Theorem 1 is met.

The formulation is compact: $O(|\mathcal{N}| \cdot |\mathbb{X}|^2)$ variables, $O(|\mathcal{N}| \cdot |\mathbb{X}|^2)$ binary clauses and $O(|\mathbb{X}|^2) \, |\mathcal{N}|$-ary clauses.

4 Accountable Decisions with Inv-NCS

In this section we describe how the theoretical and algorithmic tools described in Section 3 in order to assess the feasibility of the inverse NCS problem (see Def. 3) can be used to support a decision process. More precisely, we address the situation described in Section 1 where a committee has to assign alternatives either to the GOOD or the BAD category, and to account for this assignment. Section 4.1 addresses the first situation S1, where an audit is commissioned to check the compliance of the committee to its terms of reference, by referring to the notion of *possible* assignment. Section 4.2 addresses the second situation S2, where the committee is challenged by a stakeholder to defend a specific decision, by referring to the notion of *necessary* assignment.

4.1 Auditing Conformity

We consider the situation S1 depicted in Section 1, where an independent audit agency has to check that the decision α of the committee on on candidates $\mathbb{X}$ is compatible with NCS. We assume $\mathbb{X}^\star = \emptyset$: all the assignments should be justified together, and none should be taken for granted.

Should the burden of proof be left to the auditor, the audit procedure could require either i) full disclosure of the preference profile $\langle (\mathbb{X}, \succsim_i) \rangle_{i \in \mathcal{N}}$, and the auditor solving the NP-hard Inv-NCS problem, e.g. using a SAT solver and Corollary 2; or ii) full disclosure of the approved sets $\langle \mathcal{A}_i \rangle_{i \in \mathcal{N}}$, and the auditor solving the tractable Inv-NCS with fixed accepted sets problem as described by Proposition 1.

If we consider putting the burden of proof on the committee, Theorem 1 can be leveraged to compute and provide a certificate of feasibility for Inv-NCS(α) that involves the disclosure of less information, as illustrated below:

Example 3. *(ex. 2 cont.) If the approved sets of the committee are $\mathcal{A}_1, \ldots, \mathcal{A}_5$, then it needs to disclose information concerning three points of view in order to prove the assignment α is consistent with an approval procedure, e.g. :*

- *according to the first point of view, b is approved (and so is a which is better than b) whereas e is not (and neither is d which is worse than e), hence the procedure is able to discriminate a, b from d, e;*

- *according to the second point of view, c is approved (and so is b which is better than c) whereas d is not (and neither is f which is worse than d), hence the procedure is able to discriminate b, c from d, f;*

- *according to the fourth point of view, c is approved (and so is a which is better than c) whereas e is not (and neither is f which is worse than e), hence the procedure is able to discriminate a, c from e, f.*

The following table summarizes the points of view permitting to discriminate each pair:

		BAD		
		d	e	f
	a	1	1	4
GOOD	b	1	1	2
	c	2	4	2

This manner of arguing that a given assignment is indeed a possible outcome of an approval sorting procedure can be formalized into an *argument scheme*, an operator tying a tuple of premises – pieces of information satisfying some conditions – to a conclusion [Walton, 1996].

Definition 6 (Argument Scheme (AS1)). *We say a tuple $\langle (i_1, g_1, G_1, b_1, B_1), \ldots, (i_n, g_n, G_n, b_n, B_n) \rangle$ instantiates the argument scheme AS1 supporting the assignment α if: i) for all $k \in \{1 \ldots n\}$, $i_k \in \mathcal{N}$, $g_k \in G_k$, $\alpha(G_k) = \{\text{GOOD}\}$, $\forall g \in G_k, g \succsim_{i_k} g_k$, $b_k \in B_k$, $\alpha(B_k) = \{\text{BAD}\}$, $\forall b \in B_k, b_k \succsim_{i_k} b$ and $g_k \succ_{i_k} b_k$; and ii) $\bigcup_{k \in \{1 \ldots n\}} G_k \times B_k = \alpha^{-1}(\text{GOOD}) \times \alpha^{-1}(\text{BAD})$*

Hence, according to the point of view i_k, g_k is the least preferred alternative in the subset of GOOD alternatives G_k and it is preferred to b_k, the most preferred alternative in the subset of BAD alternatives B_k. This scheme is somewhat frugal

in the number of pairs of the profile $\langle (\mathbb{X}, \succsim_i) \rangle_{i \in \mathcal{N}}$ revealed to the auditor, as the comparisons inside $G_k \times G_k$ or $B_k \times B_k$ are not disclosed. Theorem 1 can be reworded as follows:

Corollary 3. *An assignment α is a YES instance of Inv-NCS if, and only if, there is an instance of AS1 supporting it.*

Example 4. *(Example 3 cont.) The explanations given in Example 3 instantiate AS1 as follows: $\langle (1, b, \{a, b\}, e, \{d, e\}), (2, c, \{b, c\}, d, \{d, f\}), (4, c, \{a, c\}, e, \{e, f\}) \rangle$*

The length n of an explanation instantiating the argument scheme AS1 offers an indication regarding its cognitive complexity as well as the amount of information disclosed to the auditor. Therefore, we would rather provide the shortest possible explanations, and strive to mention as few points of view as possible. Obviously, an explanation needs to reference a specific point of view at most once, so $n \leq |\mathcal{N}|$. Unfortunately, the following result shows that one might require all points of view in a complete explanation, even in situations with relatively few alternatives.

Proposition 3. *For every set of points of view $\mathcal{N}$, there exists a set of $|\mathcal{N}| + 1$ alternatives $\mathbb{X}$ and an assignment $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ for which any tuple instantiating the argument scheme AS1 and supporting α has length $|\mathcal{N}|$.*

Sketch of the Proof. The result is shown by induction on $|\mathcal{N}|$. For $|\mathcal{N}| = \{1\}$, we consider $\alpha_1 := \{(g, \text{GOOD}), (b, \text{BAD})\}$ with $g \succ_1 b$. Consider by induction an assignment α_p on p candidates $\mathbb{X}_p$ assessed on points of view $\mathcal{N} = \{1 \ldots p\}$. We introduce a new alternative z, judged as GOOD, and a new point of view $p + 1$, such that the candidates in $\mathbb{X}_p$ are indifferent on the new point of view, and z can be discriminated from b only on the new point of view. □

4.2 Justifying Individual Decisions

We now wish to justify the decision of the committee on a candidate $x \in \mathbb{X}$ (Situation S2). As we have seen in the previous section, a complete explanation of the assignment of x necessarily implies the disclosure of many information related to the other candidates, which might not be acceptable.

A possible solution is for committee to base their decision on reference cases, an assignment $\alpha^* : \mathbb{X}^* \to \{\text{GOOD}, \text{BAD}\}$, e.g. compiling past decisions that are representative of its functioning mode. In order to get rid of the influence of the other candidates, we are looking for *necessary assignments* given these reference cases.

Example 5. *(ex. 2 cont.) We consider the alternatives a, b, c, d, e, f and their assignment $\alpha^\star$ have a reference status, and we are interested in deciding on the assignment of two candidates, x, y such that:*

$$a \succ_1 f \succ_1 b \succ_1 e \succ_1 c \succ_1 y \succ_1 d \succ_1 x$$
$$e \succ_2 b \succ_2 y \succ_2 c \succ_2 d \succ_2 a \succ_2 f \succ_2 x$$
$$f \succ_3 a \succ_3 d \succ_3 b \succ_3 y \succ_3 x \succ_3 e \succ_3 c$$
$$d \succ_4 a \succ_4 c \succ_4 e \succ_4 x \succ_4 y \succ_4 f \succ_4 b$$
$$c \succ_5 y \succ_5 e \succ_5 b \succ_5 f \succ_5 x \succ_5 d \succ_5 a$$

It is not possible to represent the assignment (x, GOOD) together with the reference assignment α. Thus, x is necessarily assigned to BAD . On the contrary, both assignments (y, GOOD) and (y, BAD) can be represented together with α.

Necessary Decisions Entailed by the Jurisprudence

An explanation of the *necessity* of an assignment is intrinsically more complex than that for its *possibility*: one needs to prove that it is not possible to separate all pairs of GOOD and BAD candidates on at least one point of view. The proof relies on some deadlock that needs to be shown. Formally, this situation manifests itself in the form of an unsatisfiable boolean formula, e.g. given by Corollary 2. The unsatisfiability of the entire formula can be reduced to a $\subseteq$-minimal unsatisfiable subset of clauses (MUS), which are commonly used as certificates of infeasibility, and can also be leveraged to produce *explanations* [Junker, 2004; Besnard *et al.*, 2010; Geist and Peters, 2017]. In the case of the necessary decisions by approval sorting with a reference assignment, any MUS pinpoints a set of pairs of alternatives in $(\alpha^{-1}(\text{GOOD}) \cup \{x\}) \times \alpha^{-1}(\text{BAD})$ that cannot be discriminated simultaneously according to the points of view.

Example 6. *(ex. 5 cont.) Consider the subset of alternatives* c, d, e, f, x, *and assume* x *to be assigned to* GOOD. *Each pair in* $GB := \{(c, e), (x, d), (x, f)\}$ *needs to be discriminated from at least one point of view in* $\mathcal{N}$, *but this is not possible simultaneously: i) none of the pairs in* GB *can be discriminated neither from the first, the second nor the third point of view, as the overall* GOOD *alternative is deemed worse than the* BAD *one. ii) no more than one pair in* GB *can be discriminated according to each point of view among* $\{4, 5\}$, *and there are more pairs to discriminate than points of view.*

The pattern of deadlock illustrated by Example 6 can be generalized and formalized into an *argument scheme*, with *premises*: i) a k-tuple of pairs $\langle (g^1, b^1), \ldots, (g^k, b^k) \rangle$ of alternatives with opposite assignment, ii) a subset of points of view $B \subseteq \mathcal{N}$ with cardinality $k - 1$, such that, according to all points of view $i \notin B$, $b^j \succ_i g^j$ for all j, and, according to all points of view $i \in B$ the intervals $]b^1, g^1]_i, \ldots,]b^k, g^k]_i$ are pairwise disjoint.

Clearly, the existence of an argument instantiating the premises of this scheme is a sufficient condition for the infeasibility of representing the given assignment in the noncompensatory model, which in turn yields the *conclusion* that the candidate x is necessarily assigned to the other category.

If we assume that the cognitive burden demanded by an explanation along the lines of this argument scheme increases with the number of its premises, we derive an implicit hierarchy among the necessary decisions supported by the scheme, with a nesting $\mathcal{E}_1 \subseteq \mathcal{E}_2 \subseteq \cdots \subseteq \mathcal{E}_{|\mathcal{N}|+1}$, where $\mathcal{E}_k$ denotes the set of decisions supported by a scheme with premises referencing at most k pairs of alternatives with opposite assignment. $\mathcal{E}_1$ is exactly the set of decisions stemming from Pareto dominance, where a candidate is either at least as good as a reference alternative in the GOOD category, or at most as good as a reference alternative in the BAD category.

The question of deciding if this scheme captures a necessary condition, *i.e.* if any decision entailed by the jurisprudence can be supported by such an explanation, is left open.

Ambivalent Situations

It may happen that, for a given candidate, both assignments to GOOD and to BAD are possible. This situation is obviously all the more frequent as the reference set is small, or the number of points of view is high. In such a case, a design option would consist in constraining the decision of the committee, either favorably (e.g. following an *innocent unless proven guilty* principle) or unfavorably (e.g. following a *precautionary principle*). Another, more common, venue would give the freedom of choice to the committee. In this case, as opposed to the situation where the decision is entailed by the jurisprudence, and where the committee just needs to make obvious the link between the current case and the reference cases, the committee needs to disclose some information concerning its inner functioning. In some cases, though, Proposition 1 offers a solution that avoids a complete disclosure: suppose that, given the approved sets $\langle \mathcal{A}_i \rangle$, the candidate is approved from a coalition of points of view that is known to be insufficient (resp. sufficient), because a reference alternative is assigned to the BAD (resp. GOOD) category in a similar, or even better (resp. worse) situation than the candidate. This fortunate situation circumvents the need of discussing the particulars of the set of sufficient coalitions by referring to its upper bound $\mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ (resp. lower bound $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$).

Example 7. *(ex. 6 cont.) According to the first point of view,* y *is disapproved, as it is worse than* $c \notin \mathcal{A}_1$. *According to the third point of view,* y *is disapproved, as it is worse than* $b \notin \mathcal{A}_3$. *According to the fifth point of view,* y *is disapproved, as it is worse than* $f \notin \mathcal{A}_5$. *Furthermore, being approved according to both the second and fourth points of view is not enough to warrant access to the* GOOD *category, as illustrated by* e. *Hence,* y *is assigned to the* BAD *category.*

5 Related Work and Conclusion

In this paper we are interested in the problem of accountability of decisions issued from a noncompensatory sorting model (NCS) [Bouyssou and Marchant, 2007]. Two situations have been mainly studied. In the first one, the committee needs to justify that its decision is a possible NCS assignment. A characterization result helps to turn the existence of such assignment to finding separations of the pairs of GOOD and BAD candidates over at least one point of view, which can be formulated as a SAT problem. This allows us to generate a single argument scheme that can explain all possible NCS assignments. The second situation arises when the assignment of a new candidate is necessarily derived from jurisprudence. Thanks to the characterization result, one can also construct an argument scheme representing deadlock situations. The use of argument schemes as formal tools to convey explanation in the context of multi-criteria aiding has also been advocated in [Labreuche, 2011; Nunes *et al.*, 2014; Belahcene *et al.*, 2017].

Our solutions stem from an original take of the dual notions of *possibility* and *necessity*, often used in so-called robust optimization, decision making [Greco *et al.*, 2010] or voting contexts [Boutilier and Rosenschein, 2016] to account for incomplete information, conveying epistemic stances of skepticism or credulousness. Instead we use them to describe the leeway left to the committee in setting its expectations: the decisions taken are bound from above by possibility, described as the feasibility of the Inv-NCS problem related to

their decision, and from below by necessity, described as the infeasibility of the Inv-NCS problem simultaneously related to the reference cases and impossible assignments.

Barrot *et al.* (2013) study the problem of identifying the *possible* winners of an approval election, when votes are given but approval thresholds are unspecified. They show that determining whether a set of candidates are co-winners is NP-complete when voters have fixed (even equal) importance. Approval voting has been studied in the context of multi-winner elections [Aziz *et al.*, 2015], which may seem close to our setting: indeed, we could see the candidates ranked in GOOD as the winners. However, in our context, each candidate is ranked without consideration to the other candidates, and voters are not assumed to have equal importance.

Finally, several algorithms have been proposed to learn the parameters of a noncompensatory sorting model from observation: [Leroy *et al.*, 2011] relies on a MIP formulation, [Sobrie *et al.*, 2015] relies on a metaheuristic.

Acknowledgments

References

[Aziz *et al.*, 2015] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Simon Mackenzie, Nicholas Mattei, and Toby Walsh. Computational aspects of multi-winner approval voting. In *Proceedings of AAMAS*, pages 107–115, 2015.

[Barrot *et al.*, 2013] Nathanaël Barrot, Laurent Gourvès, Jérôme Lang, Jérôme Monnot, and Bernard Ries. Possible winners in approval voting. In *Proceedings of the third International conference on Algorithmic Decision Theory*, pages 57–70, 2013.

[Belahcene *et al.*, 2017] Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. A model for accountable ordinal sorting. In *Proceedings of the 26 International Joint Conference on Artificial Intelligence*, pages 814–820, 2017.

[Besnard *et al.*, 2010] Philippe Besnard, Éric Grégoire, Cédric Piette, and Badran Raddaoui. MUS-based generation of arguments and counter-arguments. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pages 239–244, 2010.

[Boutilier and Rosenschein, 2016] Craig Boutilier and Jeffrey S. Rosenschein. *Incomplete Information and Communication in Voting*, page 223–258. Cambridge University Press, 2016.

[Bouyssou and Marchant, 2007] Denis Bouyssou and Thierry Marchant. An axiomatic approach to noncompensatory sorting methods in MCDM, i: The case of two categories. *EJOR*, 178(1):217–245, 2007.

[Bouyssou *et al.*, 2006] Denis Bouyssou, Thierry Marchant, Marc Pirlot, Alexis Tsoukiàs, and Philippe Vincke. *Evaluation and decision models with multiple criteria: Stepping stones for the analyst*. International Series in Operations Research and Management Science. Springer, 2006.

[Doshi-Velez *et al.*, 2017] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of AI under the law: The role of explanation. *CoRR*, abs/1711.01134, 2017.

[Geist and Peters, 2017] Christian Geist and Dominik Peters. Computer-aided methods for social choice theory. In Ulle Endriss, editor, *Trends in Computational Social Choice*, chapter 13, pages 249–267. AI Access, 2017.

[Greco *et al.*, 2010] Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research*, 207(3):1455 – 1470, 2010.

[Junker, 2004] Ulrich Junker. Quickxplain: Preferred explanations and relaxations for over-constrained problems. In *Proceedings of the 19th National Conference on Artifical Intelligence*, pages 167–172, 2004.

[Kroll *et al.*, 2017] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2017.

[Labreuche, 2011] Christophe Labreuche. A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence Journal*, 175:1410–1448, 2011.

[Laslier and Sanver, 2010] Jean-François Laslier and M. Remzi Sanver. *Handbook on Approval Voting*. Studies in Choice and Welfare. Springer, Boston, 2010.

[Leroy *et al.*, 2011] Agnes Leroy, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a multiple criteria sorting method. In *Proceedings of the second International conference on Algorithmic Decision Theory*, pages 219–233, 2011.

[Mitchell, 1982] Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[Nunes *et al.*, 2014] Ingrid Nunes, Simon Miles, Michael Luck, Simone Diniz Junqueira Barbosa, and Carlos José Pereira de Lucena. Pattern-based explanation for automated decisions. In *Proceedings of 21st ECAI*, pages 669–674, 2014.

[Sobrie *et al.*, 2015] Olivier Sobrie, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a non compensatory sorting model. In *Proceedings of the fourth International Conference on Algorithmic Decision Theory*, volume 9346, pages 153–170, 2015.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2):76–99, May 2017.

[Walton, 1996] Douglas Walton. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996.

Justifying Dominating Options when Preferential Information is Incomplete

Christophe Labreuche[1] and **Nicolas Maudet**[2] and **Wassila Ouerdane**[3]

Abstract. Providing convincing explanations to accompany recommendations is a key issue in decision-aiding. In the context of decisions involving multiple criteria, the problem is made very difficult because the decision model itself may involve a complex process. In this paper, we investigate the following issue: when the preferential information provided by the user is incomplete, is there a principled way to define what is a "simple" explanation for a recommended choice? We argue first that explanations may necessitate different levels of detail. Next, we show that even when a detailed explanation is necessary, it is possible to distinguish explanations of different levels of complexity. Our results rely on an original connection we establish between the "mechanics" required to compute supporting coalitions of criteria and the simplicity of the explanation.

1 Introduction

From the first expert systems to the recent recommendation systems which flourish on commercial websites, decision-aiding has been a central concern in AI. Very soon, it has become clear that providing recommendations was only part of the challenge. Indeed, *explaining* the recommended choice(s) to the decision-maker is crucial to improve the acceptance of the recommendation [9, 3, 13], but also sometimes to allow the decision-maker to justify in turn the decision against other stakeholders. What makes an explanation "convincing" is thus highly context dependent.

In a context of movie recommendation, [7] notoriously reports that a very efficient explanation is that "this recommender system has correctly predicted 80% of the time in the past". In contexts involving more critical decisions or other users, much more detailed explanations would have to be considered [11]. Of course the ultimate nature of the explanation will depend on the underlying decision model and/or on the nature of the data provided by the user. Following [7], a useful distinction to make is among *data-based* and *process-based* explanations. To put it simply, in order to explain a recommendation, a data-based approach will focus on some key data, whereas a process-based one would make explicit (part of) the steps that lead to the decision. Both aspects are considered in this paper.

We start with a collection of partial orders over the options, as provided by different *weighted* criteria (or agents). The decision model we rely on is based on the *weighted Condorcet* principle: options are compared in a pairwise fashion, and an option a is preferred to another option b when the cumulated support that a is better than b

[1] Thales Research & Technology, 91767 Palaiseau Cedex, France, email: christophe.labreuche@thalesgroup.com
[2] LIP6, Université Paris-6, 75006 Paris Cedex 06, France, email: nicolas.maudet@lip6.fr
[3] LGI, Ecole Centrale de Paris, Chatenay Malabry, France, email: wassila.ouerdane@ecp.fr

outweighs the opposite conclusion. Our aim in this paper is to provide a principled way to produce explanations to the fact that a given set of options A is *dominating* with respect to the other options, more specifically in the sense that A constitutes a *Smith set* [4].

This decision model is specified from some *preferential information* (PI) provided during interview, related to the comparison of the options on each criterion and also on the weights of criteria. Most of the time, the PI is not sufficient to uniquely specify the model. In particular, some options may be incomparable on some criteria for the decision-maker. Moreover, the elicitation process will not result in a single value of the weight vector, but rather in a set of vectors that are compatible with the PI [6]. For instance, in the context of multi-criteria decision aid (MCDA), the decision-maker provides a few learning examples that yield constraints on the weights. In social choice, instead of assigning a weight to each party, one may only know a subset of the winning coalitions (A *winning* coalition beats its complement). Then an option is said to be necessarily preferred to another one if the first option is preferred to the second for all weight vectors that are compatible with the PI, and for all orderings of the options on the criteria that are compatible with the PI [6].

Unfortunately, when the PI is incomplete, the explanation may be quite complex, even for problems of small size, because one cannot display the value of the weights to the audience as part of an explanation. Consider the following example.

Example 1. *There are 7 options $\{a, b, c, d, e, f, g\}$ and 4 criteria $\{1, 2, 3, 4\}$. The partial orderings (noted $\succ_1, \succ_2, \succ_3, \succ_4$) of options over the 4 criteria are depicted in Figure 1. The PI regarding the importance of the criteria is composed of three items:*

- *1 together with 3 are more important than 2 and 4 together;*
- *2 and 3 together are more important than criterion 1 taken alone;*
- *4 is more important than criteria 2 and 3.*

Actually, option a is the unique dominating option. The "technical" reason is that (i) a dominates e and f on all criteria, (ii) coalition 1, 2, 3 is a winning coalition (preference of a over b), (iii) coalition 1,4 is a winning coalition (preference of a over d), (iv) coalition 1,3,4 is winning (preference of a over g), and (v) coalition 2,3,4 is a winning coalition (preference of a over c). But these reasons vary in terms of the effort required to understand them: (i) is trivial, and (ii), (iii) and (iv) are reinforcement of some statements of the PI. For instance, (ii) easily follows from the fact that 1 and 3 are already more important than 2 and 4. On the other hand, the underlying justification to (v) is more complex. How to deduce indeed from the PI, the statement that coalition 3,4 beats coalition 1,2?

We will focus on MCDA but our approach can be used in social choice in a similar way. This paper advances the state of the art by characterizing minimal complete explanations to justify dominating
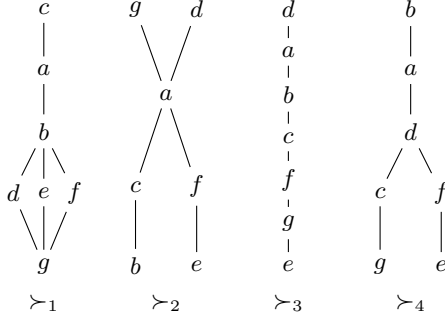
Figure 1. Partial preferences $\succ_1, \succ_2, \succ_3, \succ_4$ over the criteria 1,2,3,4.

sets in the presence of incomplete preferential information. We argue that this question calls for a process-based approach whereby comparative statements can be produced. We make precise the intuition that explanations can be of different levels of detail and complexity. Specifically, we classify explanations depending on the "operators" that were used to derive the desired statements. The major ingredient is a characterization of statements on weights that can be deduced (in terms of linear combinations) from the PI. We also show how to compute them. The remainder of this paper is as follows. Sect. 2 first details the available preferential information, the decision model, and the language used to produce explanations. We show how our approach can flexibly cater for the different degrees of accuracy that may occur within the same instance, depending whether the pairwise comparison under analysis is considered tight (Sect. 3) or obvious (Sect. 4). Sect. 5 illustrates the output as produced by our implementation. Sect. 6 discusses related work and concludes.

2 Background and basic definitions

We consider a finite set O of options and a finite set $H = \{1, \dots, m\}$ of criteria. To simplify notation, coalition $\{1, 2, 3\}$ will be noted 123.

2.1 Description of the preferential information

The decision-maker needs to provide information regarding the ranking of options O, but also regarding the relative strength of coalitions of criteria (2^H). Thus, two types of statements are considered.

Definition 1. *A preferential statement (p-statement) is of the form $[b \succ_i c]$ where $b, c \in O$ and $i \in H$, meaning that b is preferred over c on criterion i. Let $\mathcal{S}$ denote the set of all such statements.*

Definition 2. *A comparative statement (c-statement) is of the form $[I \succ J]$ where $I, J \subseteq H$ with $I \cap J = \emptyset$, meaning that the importance of the criteria in I is larger than that of the criteria in J. Let $\mathcal{V}$ denote the set of all such statements.*

It is important to remark that expressing a c-statement amounts to expressing a constraint on the feasible weight vectors attached to the criteria. Let $\mathcal{W}$ (the set of normalized weights) be the set of weights vectors $w \in [0, 1]^H$ such that $\sum_{i \in H} w_i = 1$.

We now define the operators which will make the link between the c-statements and their semantical counterpart (the weights).

Definition 3. *For a set $V \subseteq \mathcal{V}$ of c-statements, let $V^\downarrow := \{w \in \mathcal{W} \text{ s.t. } \forall [I \succ J] \in V, \ \sum_{i \in I} w_i > \sum_{i \in J} w_i\}$ be the set of weights satisfying the comparative statements V. Conversely, the set of c-statements that can be deduced from $W \subseteq \mathcal{W}$ is $W^\uparrow := \{[I \succ J] \in \mathcal{V} \text{ s.t. } \forall w \in W, \ \sum_{i \in I} w_i > \sum_{i \in J} w_i\}$. Finally, we introduce some notation:*

- *For $V \subseteq \mathcal{V}$, we set $V^{\downarrow\uparrow} = (V^\downarrow)^\uparrow$ and $\mathrm{cl}(V) := V^{\downarrow\uparrow}$.*
- *For $W \subseteq \mathcal{W}$, we set $W^{\uparrow\downarrow} = (W^\uparrow)^\downarrow$.*

Definition 4. *A PI is a pair $\langle S, V \rangle$ with $S \subseteq \mathcal{S}$ and $V \subseteq \mathcal{V}$.*

The information provided by the decision-maker is supposed to be "rational". Specifically, this means that the S part of the PI constitutes a *partial order* (reflexive, antisymmetric, transitive, but not necessarily complete), and that V is assumed to be consistent[4], in the sense that $V^\downarrow \neq \emptyset$. Note finally that the set of all linear extensions that can be obtained from S is denoted $\mathcal{S}_{\mathrm{lin}}(S)$.

Example 2 (1 ctd.). *Given the PI of Ex. 1, $V = \{[13 \succ 24], [23 \succ 1], [4 \succ 23]\}$. We have e.g. $[c \succ_1 d] \in S$, $[b \succ_2 a] \notin S$, and $\langle 0.2, 0.1, 0.15, 0.55 \rangle \notin V^\downarrow$ (violation of the first constraint).*

2.2 Description of the choice problem

Definition 5. *A set $I \subseteq H$ is called a winning coalition (w.r.t. the PI $\langle S, W \rangle$) if $\sum_{i \in I} w_i > \frac{1}{2}$ for all $w \in V^\downarrow$.*

Option b is necessarily preferred to c if whatever the weight vector compatible with V, whatever the completion of S to form total orders on each criterion, the sum of the weights of criteria supporting b is larger than the sum of the weights of criteria supporting c.

Definition 6. *For $b, c \in O$, b is necessary preferred to c given $\langle S, V \rangle$ (noted $b \succ_{S,V} c$) if*

$$\forall w \in V^\downarrow \ \forall K \in \mathcal{S}_{\mathrm{lin}}(S) \sum_{i \in H \,,\, [b \succ_i c] \in K} w_i > \sum_{i \in H \,,\, [c \succ_i b] \in K} w_i.$$

This corresponds to the *necessary preference relation* [6]. In qualitative decision models, this concept is similar to the dominance query for the CP nets [2]. An option $a \in O$ is called *weighted Condorcet winner w.r.t.* $\langle S, V \rangle$ (noted $\mathrm{WCW}_O(S, V)$) if for all $b \in O \setminus \{a\}$, $a \succ_{S,V} b$. When the WCW does not exist, it is usual to consider the Smith set (henceforth denoted by A). It is the smallest set of alternatives such that all elements of A beat all options outside this set. It is well defined and unique [4]. When a WCW exists, the Smith set is reduced to the WCW. Moreover, we set $O^\star := O \setminus A$.

There is a clear relationship between the size of A and $\langle S, V \rangle$ provided: the less informative $\langle S, V \rangle$, the more likely it is that some options cannot be compared. Specifically, the size of A will typically shrink as $\langle S, V \rangle$ gets more specified. In Example 1, e and g are incomparable because (as we shall see later) 1 and 23 are not winning coalitions given V, and e and g are incomparable on criterion 4.

This model is widely used in MCDA (note that all weights remain hidden to the user). Models not based on numerical weights also exist, but they allow less deductions to be drawn from the PI.

2.3 Description of the language for the explanation

In Example 1, we have $A = \{a\}$. When analyzing why a beats all options in $O^\star$, one notices that there are different situations. For options b, c, d, the preference of a over these options is not so trivial and deserves an adequate explanation. For option g, the case seems more clear, since a beats g on 134, and coalitions 13 and 34 are already winning coalitions. Now regarding options e, f, the dominance of a is clear since a is supported by unanimity of the criteria. Generalizing this example, it appears that dominated options can be partitioned

---

[4] In fact, many work dealing with explanations in AI address the problem of exhibiting subsets of constraints provoking an inconsistency, see *e.g.* [8].

into different classes, capturing the fact that some of them are *obviously* dominated, some are *clearly* dominated, while some others are close to a tie with some element of A. Thus, the level of detail expected by the decision maker in the produced explanation will vary.

- *unanimous* – this case occurs when an alternative b lies behind a on all criteria (technically, the option is Pareto-dominated), i.e. for all $i \in H$, $[a \succ_i b] \in S$. This requires no specific explanation.
- *large majority* – this occurs when the minimum guaranteed value of the weight of the criteria supporting a against b is larger than a threshold $\rho \in (\frac{1}{2}, 1)$ to be fixed by the designer:

$$\min_{w \in V^\downarrow} \sum_{i \in H, [a \succ_i b] \in S} w_i > \rho. \tag{1}$$

As the decision is clear-cut, the decision-maker does not need for a precise explanation.

- *weak majority* – these are the remaining cases, i.e. when the decision is not clear and a detailed explanation is required. We will focus our development mainly on this case.

The explanation process is thus as follows. For each element a in the Smith set, we denote by $O^\star_{una}[a]$, $O^\star_{large}[a]$ and $O^\star_{weak}[a]$ the set of options from $O^\star$ in the situations *unanimous*, *large majority* and *weak majority* respectively with a. The first set is easily constructed. The second will be studied at the end of the paper as we focus our analysis on the *weak majority* situation. In this case, we notice that a is a WCW of the set $O^\star_{weak}[a] \cup \{a\}$ of options (denoted by $\mathrm{WCW}_{O^\star_{weak}[a] \cup \{a\}}(S, V)$). We can thus treat each element of A separately and explain why it is a WCW of this subset of options.

3　Complete explanations for a weak majority

We turn our attention to explanations as to why $\mathrm{WCW}_{O^\star_{weak}[a] \cup \{a\}}(S, V) = \{a\}$ for some $a \in A$. We shall formally distinguish different levels of complexity required to explain c-statements in this context. This will provide a formal basis for the definition of *minimal* explanations.

3.1　Complete explanations on S and V

Following the *data-based approach* in [7], providing an explanation amounts to simplify the PI provided by the decision-maker. Accordingly, a complete explanation is a set of p-statements S' together with a set of c-statements V' such that, for any weight vector which can be deduced from V', any completion of the set of p-statements from S' yields a as a WCW. By complete, we mean that while simplifying the data, one can still prove that a is a WCW.

Definition 7. *The set of* data-based complete explanations *given* $\langle S, V \rangle$ *is:* $Ex^{\mathrm{Data}}_{S,V}(a) = \left\{ \langle S', V' \rangle \subseteq S \times V \ s.t. \ \mathrm{WCW}_{O^\star_{weak}[a] \cup \{a\}}(S', V') = \{a\} \right\}$.

We need the following definition to show that one can use a condition on the operator cl to prove that a is a WCW.

Definition 8. $P_S(a, b) := \{i \in H \ s.t. \ [a \succ_i b] \in S\}$ *and* $\mathcal{V}(S) := \{[P_S(a, b) \succ H \setminus P_S(a, b)], \ b \in O^\star_{weak}[a]\}$.

Lemma 1. $\mathrm{WCW}_{O^\star_{weak}[a] \cup \{a\}}(S', V') = \{a\}$ *iff* $\mathcal{V}(S') \subseteq \mathrm{cl}(V')$.

Proof : Let $b \in O^\star_{weak}[a]$. Let $L := \{[a \succ_i b] \ s.t. \ [a \succ_i b] \in S'\} \cup \{[b \succ_i a] \ s.t. \ [a \succ_i b] \notin S'\}$. We have $a \succ_{S', V'} b$ iff

$\sum_{i \in H, [a \succ_i b] \in L} w_i > \sum_{i \in H, [a \succ_i b] \notin L} w_i$ for all $w \in V'^\downarrow$, iff $[P_{S'}(a, b) \succ H \setminus P_{S'}(a, b)] \in V'^{\downarrow\uparrow} = \mathrm{cl}(V')$. ■

From the previous lemma, $\langle S', V' \rangle \subseteq S \times V$ is an element of $Ex^{\mathrm{Data}}_{S,V}(a)$ iff $\mathcal{V}(S') \subseteq \mathrm{cl}(V')$.

Example 3. *Consider 5 criteria and four options a, b, c, d. Assume that $V = \{[1 \succ 23], [34 \succ 15], [2 \succ 5]\}$ and $S = \{[a \succ_1 b], [a \succ_4 b], [a \succ_5 b], [a \succ_2 c], [a \succ_3 c], [a \succ_4 c], [a \succ_1 d], [a \succ_3 d], [a \succ_4 d], [b \succ_3 d]\}$. Let $V' = \{[1 \succ 23], [34 \succ 15]\}$ and $S' = S \setminus \{[b \succ_3 d]\}$. We note that $\langle S', V' \rangle \in Ex^{\mathrm{Data}}_{S,V}(a)$. In the data-based approach, $\langle S', V' \rangle$ is the minimal complete explanation in the sense of set inclusion. However, for the decision-maker, the sole knowledge of $\langle S', V' \rangle$ is not sufficient to understand why a is a WCW, i.e. why the sets of criteria $P_{S'}(a, b)$, $P_{S'}(a, c)$ and $P_{S'}(a, d)$ appearing in S' form winning coalitions. In other words, the decision maker needs to understand why $\mathcal{V}(S') = \{[145 \succ 23], [234 \succ 15], [134 \succ 15]\}$ can be deduced from V.*

Following this example, one sees that there is a major distinction between the data-based and the process-based approaches. The first one does not allow a complete traceability from the PI to the recommendations. Hence we adopt the second one in this paper regarding the c-statements. In a process-based approach, a complete explanation is a pair composed of $S' \subseteq S$ such that $\mathcal{V}(S') \subseteq \mathrm{cl}(V)$ (proving that a is a WCW), and of an explanation noted $Ex^{\mathrm{Proc}}_V(\mathcal{V}(S'))$ of why $\mathcal{V}(S')$ results from V. Ex^{Proc}_V is not further described here; it will be done in Section 3.3.

Definition 9. *The set of* process-based complete explanations *given* $\langle S, V \rangle$ *is:* $Ex^{\mathrm{Proc}}_{S,V}(a) = \left\{ \langle S', Ex^{\mathrm{Proc}}_V(\mathcal{V}(S')) \rangle : S' \subseteq S \ and \ \mathcal{V}(S') \subseteq \mathrm{cl}(V) \right\}$.

In order to be able to compute S' but also to explain $\mathcal{V}(S')$, we need to give some properties of cl and characterize $\mathrm{cl}(V)$.

3.2　cl **as closure**

From Def. 3, $\mathrm{cl}(V)$ is the set of c-statements that can be deduced from V. A natural question is whether applying cl several times adds more c-statements. We show that this is not the case. More precisely, the operator $\mathrm{cl} : \mathcal{P}(\mathcal{V}) \to \mathcal{P}(\mathcal{V})$ is a *closure*, i.e. $V \subseteq \mathrm{cl}(V)$ for all $V \subseteq \mathcal{V}$ (extensiveness), $V \subseteq V'$ implies that $\mathrm{cl}(V) \subseteq \mathrm{cl}(V')$ for all $V, V' \subseteq \mathcal{V}$ (increasingness), and $\mathrm{cl} \circ \mathrm{cl} = \mathrm{cl}$ (idempotency), as its notation suggests.

Lemma 2. *The operator* cl *is a closure.*

Proof : The following three results are clear.

$$W \subseteq W^{\uparrow\downarrow} \text{ for all } W \subseteq \mathcal{W}. \tag{2}$$
$$V \subseteq V^{\downarrow\uparrow} \text{ for all } V \subseteq \mathcal{V}. \tag{3}$$
$$\forall V, V' \in \mathcal{V}, \quad V \subseteq V' \Rightarrow V^\downarrow \supseteq V'^\downarrow. \tag{4}$$

We now give a few useful assertions.

Assertion 1. $W^\uparrow = W^{\uparrow\downarrow\uparrow}$ *for all* $W \subseteq \mathcal{W}$.

Proof : We need only to prove $W^{\uparrow\downarrow\uparrow} \subseteq W^\uparrow$ as the opposite inclusion follows from (2). Let us consider thus $[I \succ J] \in W^{\uparrow\downarrow\uparrow}$. Hence

$$\forall w \in W^{\uparrow\downarrow} \quad \sum_{i \in I} w_i > \sum_{i \in J} w_i. \tag{5}$$

Let us fix now $w \in W$. From (2), $w \in W^{\uparrow\downarrow}$. By (5), we have $\sum_{i \in I} w_i > \sum_{i \in J} w_i$. This latter relation is satisfied for all $w \in W$. Hence $[I \succ J] \in W^\uparrow$. ■

Assertion 2. $V^\downarrow = V^{\downarrow\uparrow\downarrow}$ *for all* $V \subseteq \mathcal{V}$.

Proof : Similar to that of Assertion 1. ∎

Extensiveness: Follows from (3).
Increasingness: Let $V \subseteq V'$ and $[I \succ J] \in V^{\downarrow\uparrow}$. Then for all $w \in V^\downarrow$, $\sum_{i \in I} w_i > \sum_{i \in J} w_i$. As $V'^\downarrow \subseteq V^\downarrow$ (by (4)), then for all $w \in V'^\downarrow$, $\sum_{i \in I} w_i > \sum_{i \in J} w_i$. Hence $[I \succ J] \in V'^{\downarrow\uparrow}$, and thus $V^{\downarrow\uparrow} \subseteq V'^{\downarrow\uparrow}$.

Idempotency: Follows from Assertions 1 and 2. ∎

3.3 Explanations of c-statements

The aim of this section is to construct $Ex_V^{\mathrm{Proc}}(V')$ for $V' \subseteq \mathrm{cl}(V)$. To this end, one shall explain how any element of $\mathrm{cl}(V)$ results from V. We start with a simple example.

Example 4 (2 ctd.). *From* $[4 \succ 23] \in V$*, we can deduce a fortiori that* $[14 \succ 23] \in \mathrm{cl}(V)$ *and* $[4 \succ 3] \in \mathrm{cl}(V)$.

Monotonicity generalizes the previous example in the following way:

$$[I \succ J] \in V \implies \forall I' \supseteq I \; \forall J' \subseteq J \quad [I' \succ J' \setminus I'] \in \mathrm{cl}(V) \quad (6)$$

Consider a more complex extension.

Example 5 (2 ctd.). $V^\downarrow$ *is composed of the weights* $w \in \mathcal{W}$ *satisfying* $w_1 + w_3 > w_2 + w_4$, $w_2 + w_3 > w_1$ *and* $w_4 > w_2 + w_3$.

New constraints can be derived by linear combinations of these constraints. For instance, the constraint $w_3 + w_4 > w_1 + w_2$ *results from the summation of constraint* $w_1 + w_3 > w_2 + w_4$ *with two times the constraints* $w_2 + w_3 > w_1$ *and* $w_4 > w_2 + w_3$.

The next proposition shows that the intuition of Example 5 holds in the general case: all c-statements that can be deduced from V results from linear combinations (with integer coefficients) of the constraints in V and of the constraints on the sign of the weights.

Proposition 1. $[I \succ J] \in \mathrm{cl}(V)$ *iff the following ILP is feasible:*

Find non-negative integers $\{\alpha_{E,F}\}_{[E \succ F] \in V}, \{\beta_i\}_{i \in H}, \gamma$

minimizing $\sum_{[E \succ F] \in V} \alpha_{E,F} + \sum_{i \in H} \beta_i + \gamma$ *such that*

$$\sum_{[E \succ F] \in V} \alpha_{E,F} \geq 1 \qquad (7)$$

$$\beta_i + \sum_{[E \succ F] \in V, \, E \ni i} \alpha_{E,F} - \sum_{[E \succ F] \in V, \, F \ni i} \alpha_{E,F} = \begin{cases} \gamma & \text{if } i \in I \\ -\gamma & \text{if } i \in J \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

for all $i \in H$.

Proof : The normalization condition of the weights can be removed since we analyse the completion among comparative statements. Let $U := \{w \in \mathbb{R}_+^m : \forall [E \succ F] \in V, \sum_{i \in E} w_i > \sum_{i \in F} w_i\}$. Let us consider $[I \succ J] \in \mathrm{cl}(V)$. Hence for all $w \in U$, $\sum_{i \in I} w_i > \sum_{i \in J} w_i$. This means that $U' := \{w \in U, \sum_{i \in E} w_i \leq \sum_{i \in F} w_i\} = \emptyset$. Hence the linear constraints in U' are inconsistent. From Motzkin's theorem [12, pages 28-29], there exists non-negative integers $\alpha_{E,F}, \gamma$ and β_i with at least one coefficient corresponding to the strict inequalities (i.e. at least one $\alpha_{E,F}$ non-zero – see (7)) such that the coefficients in front of each w_i in the following expression are equal to zero

$$\sum_{[E \succ F] \in V} \alpha_{E,F} \left(\sum_{i \in E} w_i - \sum_{i \in F} w_i \right) + \gamma \left(\sum_{i \in J} w_i - \sum_{i \in I} w_i \right) + \sum_{i \in H} \beta_i \, w_i.$$

Hence (8) is fulfilled for all $i \in H$. ∎

The previous proposition is very important. It provides a characterization of $\mathrm{cl}(V)$. It shows precisely how constraint $[I \succ J]$ is derived from V and the sign of the weights. The values $\alpha_{E,F}$ and β_k are the coefficients that are multiplied by the constraints $\sum_{i \in E} w_i > \sum_{i \in F} w_i$ and $w_k \geq 0$ respectively. The summation yields the constraint $\gamma \times \sum_{i \in I} w_i > \gamma \times \sum_{i \in J} w_i$.

If the coefficients α, β, γ satisfy (7) and (8), multiplying these coefficients by any positive integer also verify the constraints. The use of the minimization functional in the ILP ensures that we obtain the smallest values of the coefficients and thus the simplest explanation.

Definition 10. *The complete explanation* $Ex_V^{\mathrm{Proc}}([I \succ J])$ *of the c-statement* $[I \succ J] \in \mathrm{cl}(V)$ *is* $\langle (\{\alpha_{E,F}\}_{[E \succ F] \in V}, \{\beta_i\}_{i \in H}, \gamma), [I \succ J] \rangle$.
For $V' \subseteq \mathrm{cl}(V)$, $Ex_V^{\mathrm{Proc}}(V') := \cup_{[I \succ J] \in V'} Ex_V^{\mathrm{Proc}}([I \succ J])$.

Example 6 (5 ctd.). *With* $\alpha_{13,24} = 1$, $\alpha_{23,1} = 2$, $\alpha_{4,23} = 1$, $\beta_i = 0$ *for all* $i \in H$ *and* $\gamma = 1$, *we obtain* $[34 \succ 12] \in \mathrm{cl}(V)$.
Moreover, $\mathrm{cl}(V) = \{[123 \succ 4], [124 \succ 3], [134 \succ 2], [13 \succ 24], [13 \succ 2], [14 \succ 23], [14 \succ 2], [1 \succ 2], [13 \succ 4], [14 \succ 3], [234 \succ 1], [23 \succ 1], [24 \succ 1], [34 \succ 12], [34 \succ 1], [4 \succ 1], [24 \succ 3], [34 \succ 2], [3 \succ 2], [4 \succ 23], [4 \succ 2], [4 \succ 3]\}$.

3.4 Complexity levels in explaining c-statements

In Example 6, let us consider four particular elements of $\mathrm{cl}(V)$, $[23 \succ 1]$, $[4 \succ 3]$, $[4 \succ 1]$ and $[34 \succ 12]$. The difficulty of justifying these four statements from V is not the same. Indeed, the first statement $[23 \succ 1]$ is directly contained in V so that there is no underlying complexity for the user. The second statement $[4 \succ 3]$ is directly obtained from $[4 \succ 23] \in V$ using a monotonicity argument (see Example 4). The third statement $[4 \succ 1]$ results from the summation of the two relations $w_2 + w_3 > w_1$ and $w_4 > w_2 + w_3$ of $V^\downarrow$. Lastly, as we already noticed in Example 5, the last statement $[34 \succ 12]$ is more complex to obtain. The arguments that we use to justify a statement from $\mathrm{cl}(V)$, going from the first statement to the fourth one are of increasing complexity.

It seems thus natural to decompose $\mathrm{cl}(V)$ into four nested sets. The first set $\mathrm{cl}_0(V) := V$ is the c-statements contained in the PI. The second set $\mathrm{cl}_1(V)$ is composed of the elements of $\mathrm{cl}(V)$ that can be deduced from V only using monotonicity condition (see (6)). This corresponds to the case where, in Proposition 1, all α coefficients are equal to 0, except one that is equal to 1. The third set $\mathrm{cl}_2(V)$ is composed of the elements of $\mathrm{cl}(V)$ that can be deduced from V only using summation and monotonicity conditions. This corresponds to the case when the α coefficients are either equal to 0 or 1. Finally, $\mathrm{cl}_3(V) = \mathrm{cl}(V)$. The set $\mathrm{cl}(V)$ is partitioned in the following way.

Definition 11. $\Delta_0 = \mathrm{cl}_0(V)$, $\Delta_j = \mathrm{cl}_j(V) \setminus \mathrm{cl}_{j-1}(V)$ *for* $j \in \{1, 2, 3\}$.

Example 7 (6 ctd.). *We have* $\Delta_0 = \{[13 \succ 24], [23 \succ 1], [4 \succ 23]\}$, $\Delta_1 = \{[123 \succ 4], [134 \succ 2], [13 \succ 2], [13 \succ 4], [234 \succ 1], [124 \succ 3], [14 \succ 23], [14 \succ 2], [14 \succ 3], [24 \succ 3], [34 \succ 2], [4 \succ 2], [4 \succ 3]\}$, $\Delta_2 = \{[24 \succ 1], [34 \succ 1], [4 \succ 1], [1 \succ 2], [3 \succ 2]\}$ *and* $\Delta_3 = \{[34 \succ 12]\}$.

The sets $\Delta_0, \Delta_1, \Delta_2, \Delta_3$ are of increasing complexity. When comparing two sets $V', V'' \subseteq \mathrm{cl}(V)$, we prefer the set that has the smallest number of elements in Δ_3. In case of equality, we prefer the one that has the smallest number of elements in Δ_2. And so on. The

following ordering $\rhd_\mathcal{V}$ depicts the complexity of understanding why a set of c-statements derives from V.

Definition 12. *For* $V', V'' \subseteq \mathrm{cl}(V)$, $V' \rhd_\mathcal{V} V''$ *iff* $(|V' \cap \Delta_3|, |V' \cap \Delta_2|, |V' \cap \Delta_1|, |V' \cap \Delta_0|) \succ_{\mathrm{lex}} (|V'' \cap \Delta_3|, |V'' \cap \Delta_2|, |V'' \cap \Delta_1|, |V'' \cap \Delta_0|)$, *where* $\succ_{\mathrm{lex}}$ *is the lexicographic ordering.* $(x_1, x_2, x_3, x_4) \succ_{\mathrm{lex}} (y_1, y_2, y_3, y_4)$ *if there exists* $i \in \{1, 2, 3, 4\}$ *such that* $x_i > y_i$ *and* $x_j = y_j$ *for all* $j \in \{1, 2, 3, 4\}$ *with* $j < i$.

We notice the elements of S and of Δ_0 are of the same complexity since they are both elements of the PI. The number of elements of p-statements is added to the number of c-statements that belong to Δ_0.

Definition 13. *Let* $\langle S', Ex_V^{\mathrm{Proc}}(\mathcal{V}(S'))\rangle, \langle S'', Ex_V^{\mathrm{Proc}}(\mathcal{V}(S''))\rangle \in Ex_{S,V}^{\mathrm{Proc}}(a)$. *The complexity of* S' *is* $comp(S') := (|\mathcal{V}(S') \cap \Delta_3|, |\mathcal{V}(S') \cap \Delta_2|, |\mathcal{V}(S') \cap \Delta_1|, |\mathcal{V}(S') \cap \Delta_0| + |S'|)$. *We define the order* $\rhd$ *(over* $Ex_{S,V}^{\mathrm{Proc}}(a)$*) by* $\langle S', Ex_V^{\mathrm{Proc}}(\mathcal{V}(S'))\rangle \rhd \langle S'', Ex_V^{\mathrm{Proc}}(\mathcal{V}(S''))\rangle$ *iff* $comp(S') \succ_{\mathrm{lex}} comp(S'')$.

3.5 Determination of the minimal explanations

In order to compute the minimal explanations in the sense of $\rhd$, one may proceed in three steps: (S1) determines all elements of $\mathrm{cl}_1(V)$, $\mathrm{cl}_2(V)$ and $\mathrm{cl}_3(V)$; (S2) identifies all elements of $Ex_S^{\mathrm{Proc}}(a) := \{S' \subseteq S : \mathcal{V}(S') \subseteq \mathrm{cl}(V)\}$; (S3) determines the elements S' in $Ex_S^{\mathrm{Proc}}(a)$ such that $\langle S', Ex_V^{\mathrm{Proc}}(\mathcal{V}(S'))\rangle$ is minimal in the sense of $\rhd$. To determine whether $[I \succ J] \in \mathrm{cl}(V)$, it suffices to perform an ILP (by Proposition 1). Step (S1) requires to compute all elements of $\mathrm{cl}_1(V)$, $\mathrm{cl}_2(V)$ and $\mathrm{cl}_3(V)$, which is not necessary in step (S2) and might be time consuming. Hence, we propose to perform steps (S1) and (S2) at the same time, determining the belonging to the closure only when required.

To this end, the following variables are introduced in the algorithm: $\mathcal{Cl}_1, \mathcal{Cl}_2, \mathcal{Cl}_3$ correspond to the elements of $\mathrm{cl}_1(V)$, $\mathrm{cl}_2(V)$, $\mathrm{cl}_3(V)$ that are useful in the analysis; $\mathcal{NCl}$ are c-statements that are not in $\mathrm{cl}(V)$ and $ExCl$ contains the explanations of $\mathcal{Cl}_1, \mathcal{Cl}_2, \mathcal{Cl}_3$. We start by initializing these variables

$$\mathcal{Cl}_0 = V \ , \ \mathcal{Cl}_1 \leftarrow \emptyset \ , \ \mathcal{Cl}_2 \leftarrow \emptyset \ , \ \mathcal{Cl}_3 \leftarrow \emptyset \ , \ \mathcal{NCl} \leftarrow \emptyset \ , \ ExCl \leftarrow \emptyset.$$

The next algorithm checks whether a c-statements belongs to $\mathrm{cl}(V)$.

Algorithm 1. **Function** isInClosure(I) *returns a boolean saying whether* $[I \succ H \setminus I] \in \mathrm{cl}(V)$, *and updates* $\mathcal{Cl}_1, \mathcal{Cl}_2, \mathcal{Cl}_3, \mathcal{NCl}, ExCl$:

> **If** $[I \succ H \setminus I] \in \mathcal{Cl}_0 \cup \mathcal{Cl}_1 \cup \mathcal{Cl}_2 \cup \mathcal{Cl}_3$ **then return true**;
> **If** $[I \succ H \setminus I] \in \mathcal{NCl}$ **then return false**;
> **Launch** the ILP of Proposition 1 on $[I \succ H \setminus I]$;
> **If** the ILP is infeasible **then**
> > $\mathcal{NCl} \leftarrow \mathcal{NCl} \cup \{[I \succ H \setminus I]\}$; **return false**;
>
> $ExCl \leftarrow ExCl \cup \{\langle(\{\alpha_{E,F}\}_{[E \succ F] \in V}, \{\beta_i\}_{i \in H}, \gamma), [I \succ H \setminus I]\rangle\}$;
> **If** $\sum_{[E \succ F] \in V} \alpha_{E,F} = 1$ **then return true**;
> **If** $\alpha_{E,F} \in \{0, 1\}$ for all $[E \succ F] \in V$ **then**
> > $\mathcal{Cl}_2 \leftarrow \mathcal{Cl}_2 \cup \{[I \succ H \setminus I]\}$; **return true**;
>
> $\mathcal{Cl}_3 \leftarrow \mathcal{Cl}_3 \cup \{[I \succ H \setminus I]\}$; **return true**;

The main algorithm is now described. Firstly, it computes $\mathcal{P}_S(a, b) := \{I \subseteq P_S(a, b) \text{ s.t. } [I \succ H \setminus I] \in \mathrm{cl}(V)\}$. Then steps (S2) and (S3) are performed.

Algorithm 2. **Function** bestExplanations(a) *(for* $a \in A$*) computes the elements of* $Ex_{S,V}^{\mathrm{Proc}}(a)$ *that are minimal w.r.t.* $\rhd$:

> **For all** $b \in O_{weak}^\star[a]$ **do**
> > $\mathcal{P}_S(a, b) \leftarrow \emptyset$;
> > **For all** $I \subseteq P_S(a, b)$ **do**
> > > **If** isInClosure(I) =**true then** $\mathcal{P}_S(a, b) \leftarrow \mathcal{P}_S(a, b) \cup \{I\}$;

> $Ex_S^{\mathrm{Proc}}(a) \leftarrow \Big\{\{[a \succ_i b]\ ,\ i \in I_b \text{ and } b \in O_{weak}^\star[a]\}$
> > $\text{for all } I_b \in \mathcal{P}_S(a, b), b \in O_{weak}^\star[a]\Big\}$;
>
> $E \leftarrow \Big\{\langle S', Ex_V^{\mathrm{Proc}}(\mathcal{V}(S'))\rangle\ ,\ S' \in Ex_S^{\mathrm{Proc}}(a)\Big\}$;
> > *(where* Ex_V^{Proc} *is stored in* $ExCl$*)*
> **return** *the minimal elements of* E *w.r.t.* $\rhd$;

Note that in the first loop in Algorithm 2, when isInClosure(I) returns *false*, we need not explore any subset of I (they cannot belong to $\mathcal{P}_S(a, b)$). This treatment is clearly exponential in the number of criteria in the worst case. In Algo. 2, the number of calls to an ILP is at worse 2^m (where m is the number of criteria) and the total number of calls of isInClosure is at worse $|O_{weak}^\star[a]| \times 2^m$. Moreover, the cardinality of $Ex_S^{\mathrm{Proc}}(a)$ is at worse $2^{m \times |O_{weak}^\star[a]|}$. In practice, it might be much less – see Example 8 below. Moreover our approach only computes ILPs when required.

Example 8 (1 ctd.). *In the comparison of a with b, c, d, the number of ILP that are solved is* 10 *(instead of* 16 *in the worse case), and* isInClosure *is called* 15 *times (instead of* 48 *in the worse case). Moreover,* $|Ex_S^{\mathrm{Proc}}(a)| = 4$ *(instead of* 4096 *in the worse case) – see Section 5 for the details.*

4 Complete explanation for a large majority

We have seen at the beginning of the paper, that the *large majority* situation applies when condition (1) is satisfied.

Definition 14. $V_\rho^\Updownarrow = \Big\{I \subseteq H \text{ s.t. } \forall w \in V^\downarrow \sum_{i \in I} w_i > \rho\Big\}.$

This section is concerned with the identification of the coalitions in $V_\rho^\Updownarrow$ with the associated explanation. One first notes that if $I \in V_\rho^\Updownarrow$ then necessarily $[I \succ H \setminus I] \in \mathrm{cl}(V)$ as $\rho > \frac{1}{2}$.

Example 9. *Assume that* $V = \{[1 \succ 2], [2 \succ 3], [3 \succ 4], [4 \succ 5]\}$, *with* $m = 5$. *One can easily show that* $w_1 + w_2 + w_3 > \frac{3}{5}$ *for all* $w \in V^\downarrow$. *Hence coalition* $123 \in V_\rho^\Updownarrow$ *if ρ does not exceed* $\frac{3}{5}$.

We characterize the elements of $V_\rho^\Updownarrow$.

Proposition 2. $I \in V_\rho^\Updownarrow$ *iff the following ILP is feasible:*

> *Find* $\{\alpha_{E,F}\}_{[E \succ F] \in V} \in \mathbb{N}_+^V, \{\beta_i\}_{i \in H} \in \mathbb{N}_+^H, \delta \in \mathbb{N}, \gamma \in \mathbb{N}_+$
> *minimizing* $\sum_{[E \succ F] \in V} \alpha_{E,F} + \sum_{i \in H} \beta_i + \delta + \gamma$ *such that*

$$\sum_{[E \succ F] \in V} \alpha_{E,F} \geq 1 \tag{9}$$

$$\beta_i + \delta + \sum_{[E \succ F] \in V ,\ E \ni i} \alpha_{E,F} - \sum_{[E \succ F] \in V ,\ F \ni i} \alpha_{E,F} = \left\{ \begin{array}{l} \gamma \text{ if } i \in I \\ 0 \text{ otherwise} \end{array} \right.$$

$$\text{for all } i \in H, \text{ and} \tag{10}$$

$$\delta \geq \gamma \rho \tag{11}$$

Proof (Sketch): Similar to that of Proposition 1. The normalization condition must be considered due to the presence of non-zero right hand side in (1). The inequality (11) follows from this. ∎

The values $\alpha_{E,F}, \beta_k$ and δ are the coefficients that are multiplied by the constraints $\sum_{i \in E} w_i > \sum_{i \in F} w_i$, $w_k \geq 0$ and $\sum_{i \in H} w_i = 1$ respectively.

Example 10 (Example 9 cont.). *For* $\frac{1}{2} < \rho \leq \frac{3}{5}$, $123 \in V_\rho^\Updownarrow$ *results from the coefficients* $\alpha_{1,2} = 2, \alpha_{2,3} = 4$, $\alpha_{3,4} = 6$, $\alpha_{4,5} = 3$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, $\delta = 3$ *and* $\gamma = 5$. *More generally,* $V_\rho^\Updownarrow = \{123, 124, 1234, 1235, 1245, 12345\}$.

The generation of explanation from the coefficients in Prop. 2 can be done as for Prop. 1, based on α, β, δ and γ.

5 Output of the results

To wrap up, for all $a \in A$, $O_{una}^\star[a] = \{b \in O^\star, P_S(a,b) = H\}$, $O_{large}^\star[a] = \{b \in O^\star \setminus O_{una}^\star[a], P_S(a,b) \in V_\rho^\Updownarrow\}$ and $O_{weak}^\star[a]$ are the remaining elements of $O^\star$. ILP is used to compute the coefficients appearing in Propositions 1 and 2. The minimal explanations for $O_{weak}^\star[a]$ are obtained thanks to Algorithm 2. As for the closure, the elements of $V_\rho^\Updownarrow$ are computed only when required, that is only for coalitions $P_S(a,b)$.

We conclude the paper by considering again the Example 1, to illustrate how our approach (implemented in JAVA) outputs the results. We recall that the Smith set is a. The explanation generated is as follows:

- a is better than e and f by *unanimity* of the criteria;
- a is better than g on a *large majority*.

By default, the system shall not give any further detail, since the case is deemed clear enough not to require any further justification. Upon request (*why?*) of the decision-maker however, the algorithm may provide the following explanation.

- In fact, the large majority is 134, and $134 \in V_\rho^\Updownarrow$, with $\rho = 0.7$ and the coefficients $\alpha_{13,24} = 1$ (for $[13 \succ 24]$), $\alpha_{4,23} = 2$ (for $[4 \succ 23]$), $\beta_3 = 2$ (for $w_3 \geq 0$), $\delta = 3$ (for $w_1 + w_2 + w_3 + w_4 = 1$), $\gamma = 4$ (for $134 \in V_\rho^\Updownarrow$), and all other coefficients are zero.

Concerning the comparison of a with b, c and d, we apply Algorithm 2 described in Section 3.5. In particular, $Ex_S^{\text{Proc}}(a) = \{S_1, S_2, S_3, S_4\}$, where $S_1 = \{[a \succ_1 b], [a \succ_3 b], [a \succ_3 c], [a \succ_4 c], [a \succ_1 d], [a \succ_4 d]\}$, $S_2 = S_1 \cup \{[a \succ_2 b]\}$, $S_3 = S_1 \cup \{[a \succ_2 c]\}$ and $S_4 = S_1 \cup \{[a \succ_2 b], [a \succ_2 c]\}$. At first sight, S_1 seems the simplest set and S_4 the most complex one. This intuition is defected. By Def. 13 and Ex. 7, $comp(S_1) = (1, 0, 1, 7)$ as $\mathcal{V}(S_1)$ is composed of $[13 \succ 24] \in \Delta_0$ (comparison of a and b), $[34 \succ 12] \in \Delta_3$ (comparison of a and c) and $[14 \succ 23] \in \Delta_1$ (comparison of a and d), and S_1 is composed of 6 statements. Likewise, $comp(S_2) = (1, 0, 2, 7)$, $comp(S_3) = (0, 0, 2, 8)$, $comp(S_4) = (0, 0, 3, 8)$. Comparing S_1 and S_3, it is apparent that simplifying over the p-statements might result in a much more complex explanation regarding the c-statements. Hence the minimal element of $Ex_{S,V}^{\text{Proc}}(a)$ w.r.t. $\rhd$ is $\langle S_3, Ex_V^{\text{Proc}}(\mathcal{V}(S_3)) \rangle$. The latter takes the following form:

- a is better than b on the *weak majority* $13 \in \Delta_0$;
- a is better than c on the *weak majority* $234 \in \Delta_1$ such that $\alpha_{23,1} = 1$ (for $[23 \succ 1]$), $\beta_4 = 1$ (for $w_4 \geq 0$) and $\gamma = 1$;
- a is better than d on the *weak majority* $14 \in \Delta_1$ such that $\alpha_{4,23} = 1$ (for $[4 \succ 23]$), $\beta_1 = 1$ (for $w_1 \geq 0$) and $\gamma = 1$.

The previous explanation is very detailed. For a user who does not require such level of traceability (e.g. a user with a shallower understanding of the decision process), it is possible to hide the coefficients α, β and γ, by just mentioning the set statements that yield another one. We emphasize that we did not explore yet the natural language issues that occur here: the way to present and organize the same (content-wise) explanation may clearly affect the way it is perceived [3]. We leave this for further research.

6 Related work and Conclusion

In the domain of recommender systems, the issue of explanation has motivated a huge amount of studies. In their recent taxonomy proposal, [5] distinguish three distinctive features to classify generated explanations: the *reasoning model* (whether the explanation disclose, even partially, the decision model), the *recommendation paradigm* (the type of decision model), and the *information categories* which are used when the explanation is generated (more specifically, whether they use the user's model, whether they refer to the recommended item and/or to the alternative options). Our approach would be classified as follows: *white box*, *knowledge-based*, and using the *three* categories (in our case: user's rankings, and referring to both the recommended item and the dominated ones). A distinctive feature of our approach lies on the decision model used, taken together with the fact that the PI may be largely incomplete. In this context, the precise weights attached to attributes cannot be exhibited, and the challenge is to provide convincing (complete) explanations despite this constraint.

We also observe that there is at least a syntactic similarity with argumentation theories. For instance, in Definition 10, an explanation is a pair $\langle C, [I \succ J] \rangle$, where C is minimal and is the support of the explanation and $[I \succ J]$ is the conclusion. This may be seen as an argument, a pair $\langle H, h \rangle$ where h is the conclusion, H is a minimal consistent subset of the knowledge base that entails h [1]. However we emphasize again that in our context we are looking for proofs, whereas arguments support non-monotonic inferences. Under this more argumentative perspective, [10] puts forward the idea of having different levels of explanations for multi-attribute preference models.

A study of complete explanations for the same type of preference model can be found in [11]. The main difference is that completeness of the PI (for both weights and rankings) is assumed in [11]. In this context, a sentence like *"the weight of criterion 2 is 0.3"* may stand as a valid justification. Instead, this paper investigates the explanation of the importance of coalitions of criteria.

Acknowledgements

REFERENCES

[1] P. Besnard and A. Hunter, *Elements of argumentation*, MIT Press, 2008.

[2] C. Boutilier, R. Brafman, C. Domshlak, H. Hoos, and D. Poole, 'CP-nets: a tool for representing and reasoning with conditional Ceteris Paribus preference statements.', *JAIR*, **21**, 135–191, (2004).

[3] G. Carenini and J.D. Moore, 'Generating and evaluating evaluative arguments', *AIJ*, **170**, 925–952, (2006).

[4] P. Fishburn, 'Condorcet social choice functions', *SIAM Journal on Applied Mathematics*, **33**(3), 469–489, (1977).

[5] G. Friedrich and M. Zanker, 'A taxonomy for generating explanations in recommender systems', *AI Magazine*, **32**(3), 90–98, (2011).

[6] S. Greco, R. Słowinski, J. Figueira, and V. Mousseau, 'Robust ordinal regression', in *Trends in Multiple Criteria Decision Analysis*, 241–284, Springer Verlag, (2010).

[7] J. L. Herlocker, J. A. Konstan, and J. Riedl, 'Explaining collaborative filtering recommendations', in *CSCW*, pp. 241–250, (2000).

[8] U. Junker, 'Quickxplain: preferred explanations and relaxations for over-constrained problems', in *Proceedings of AAAI'04*, pp. 167–172, San Jose, Calfornia, USA, (2004).

[9] D.A. Klein, *Decision analytic intelligent systems: automated explanation and knowledge acquisition*, Lawrence Erlbaum Associates, 1994.

[10] Ch. Labreuche, 'A general framework for explaining the results of a multi-attribute preference model', *AIJ*, **175**, 1410–1448, (2011).

[11] Ch. Labreuche, N. Maudet, and W. Ouerdane, 'Minimal and complete explanations for critical multi-attribute decisions', in *Algorithmic Decision Theory (ADT)*, pp. 121–134, Piscataway, NJ, USA, (2011).

[12] O.L. Mangasarian, *Nonlinear Programming*, McGraw-Hill Book Comp., 1969.

[13] D. Mcsherry, 'Explanation in recommender systems', *AIR*, **24**, 179–197, (2005).

Minimal and Complete Explanations for Critical Multi-attribute Decisions

Christophe Labreuche[1], Nicolas Maudet[2], and Wassila Ouerdane[3]

[1] Thales Research & Technology
91767 Palaiseau Cedex, France
christophe.labreuche@thalesgroup.com
[2] LAMSADE, Université Paris-Dauphine
Paris 75775 Cedex 16, France
maudet@lamsade.dauphine.fr
[3] Ecole Centrale de Paris
Chatenay Malabry, France
wassila.ouerdane@ecp.fr

Abstract. The ability to provide explanations along with recommended decisions to the user is a key feature of decision-aiding tools. We address the question of providing minimal and complete explanations, a problem relevant in critical situations where the stakes are very high. More specifically, we are after explanations with minimal cost supporting the fact that a choice is the weighted Condorcet winner in a multi-attribute problem. We introduce different languages for explanation, and investigate the problem of producing minimal explanations with such languages.

1 Introduction

The ability to provide explanations along with recommended decisions to the user is a key feature of decision-aiding tools [1,2]. Early work on expert systems already identified it as one of the main challenge to be addressed [3], and the recent works on recommender systems face the same issue, see *e.g.* [4]. Roughly speaking, the aim is to increase the user's acceptance of the recommended choice, by providing supporting evidence that this choice is justified.

One of the difficulties of this question lies on the fact that the relevant concept of an explanation may be different, depending on the problem at hand and on the targeted audience. The objectives of the explanations provided by an online recommender system are not necessarily the same as the ones of a pedagogical tool. To better situate our approach, we emphasize two important distinctive dimensions:

- *data vs. process*—following [5], we first distinguish explanations that are based on the *data* and explanations that are based on the *process*. Explanations based on the data typically focus on a "relevant" subset of the available data, whereas those based on the process make explicit (part of) the mathematical model underlying the decision.

– *complete vs. incomplete explanations*—as opposed to incomplete explanations, complete explanations support the decision unambiguously, they can be seen as proofs supporting the claim that the recommended decision is indeed the best one. This is the case for instance in critical situations (*e.g.* involving safety) where the stakes are very high.

In this paper we shall concentrate on complete explanations based on the data, in the context of decisions involving multiple attributes from which, associating a preference model, we obtain *criteria* upon which options can be compared. Specifically, we investigate the problem of providing simple but complete explanations to the fact that a given option is a weighted Condorcet winner (WCW). An option is a WCW if it beats any other options in pairwise comparison, considering the relative weights of the different criteria. Unfortunately, a WCW may not necessarily exists. We focus on this case because (i) when a WCW exists it is the unique and uncontroversial decision to be taken, (ii) when it does not many decision models can be seen as "approximating" it, and (iii) the so-called outranking methods (based on the Condorcet method) are widely used in multi-criteria decision aiding, (iv) even though the decision itself is simple, providing a *minimal* explanation may not be.

In this paper we assume that the problem involves two types of *preferential information* (PI): preferential information regarding the importance of the criteria, and preferential information regarding the ranking of the different options.

To get an intuitive understanding of the problem, consider the following example.

Example 1. There are 6 options $\{a, b, c, d, e, f\}$ and 5 criteria $\{1, \cdots, 5\}$ with respective weights as indicated in the following table. The (full) orderings of options must be read from top (first rank) to bottom (last rank).

criteria	1	2	3	4	5
weights	0.32	0.22	0.20	0.13	0.13
ranking	c	b	f	d	e
	a	a	e	f	b
	e	f	a	b	d
	d	e	c	a	f
	b	d	d	c	a
	f	c	b	e	c

In this example, the WCW is a. However this option does not come out as an obvious winner, hence the need for an explanation. Of course a possible explanation is always to explicitly exhibit the computations of every comparison, but even for moderate number of options this may be tedious. Thus, we are seeking explanations that are minimal, in a sense that we shall define precisely below. What is crucial at this point is to see that such a notion will of course be dependent on the language that we have at our disposal to produce explanations. A tentative "natural" explanation would be as follows:

"First consider criteria 1 and 2, a is ranked higher than e, d, and f in both, so is certainly better. Then, a is preferred over b on criteria 1 and

> 3 (which is almost as important as criterion 2). Finally, it is true that c is better than a on the most important criteria, but a is better than c on all the other criteria, which together are more important."

The aim of this paper is not to produce such natural language explanation, but to provide the theoretical background upon which such explanations can later be generated.

This abstract example may be instantiated in the following situations. In the first one, a decision-maker presents a choice recommendation regarding a massive investment before funding agency. The decision was based on a multi-criteria analysis during which criteria and preferences were elicited. In the second one, a committee (where members have different voting weights) just proceeded to a vote on a critical issue, and the chairman is now to explain why a given option was chosen as a result. The reason why we take these two concrete examples is that beyond their obvious similarity (members of the committee play the role of the criteria in the funding example), they share the necessity to produce a complete explanation. The type of explanation we seek for is relevant when the voters (for the committee example) are not anonymous, which is often the case in committee.

The remainder of this paper is as follows. In the next section, we provide the necessary background notions, and introduce in particular the languages we shall use for formulating explanations. Section 3 defines minimal complete explanations. Section 4 and Section 5 deal with languages allowing to express the preferences on the rankings of options only, starting with the language allowing basic statements, then discussing a more refined language allowing to "factor" statements. Finally, Section 6 discusses connections to related works, in particular argumentation theory.

2   Background and Basic Definitions

2.1   Description of the Choice Problem

We assume a finite set of options O, and a finite set of criteria $H = \{1, 2, \ldots, m\}$. The options in O are compared thanks to a weighted majority model based on some preferential information (PI) composed of preferences and weights. Preferences are linear orders, that is, complete rankings of the options in O, and $a \succ_i b$ stands for the fact that a is strictly preferred over b on criterion i. Weights are assigned to criteria, and W_i stands for the weight of criterion i. Furthermore, they are normalized in the sense that they sum up to 1. An instance of the choice problem, denoted by ρ, is given by the full specification of this PI. The decision model over O given ρ is defined by $b \succ_\rho c$ iff $\sum_{b \succ_i c} W_i > \sum_{c \succ_i b} W_i$.

Definition 1. *An option $a \in O$ is called* weighted Condorcet winner w.r.t. ρ *(noted* $\mathrm{WCW}(\rho)$*) if for all $b \in O^\star := O \setminus \{a\}$, $a \succ_\rho b$.*

We shall also assume throughout this paper the existence of a weighted Condorcet winner labeled $a \in O$.

2.2   Description of the Language for the Explanation

Following the example in the introduction, the simplest language on the partial preferences is composed of terms of the form $[i : b \succ c]$, with $i \in H$ and $b, c \in O$, meaning that b is strictly preferred to c on criterion i. Such terms are called *basic preference statements*. In order to reduce the length of the explanation, they can also be factored into terms of the form $[I : b \succ P]$, with $I \subseteq H$, $b \in O$ and $P \subseteq O \setminus \{b\}$, meaning that b is strictly preferred to all options in P on all criteria in I. Such terms are called *factored preference statements*. The set of all subsets of basic preference statements (resp. factored preference statements) that correspond to a total order over O on each criterion is denoted by $\mathcal{S}$ (resp. $\widehat{\mathcal{S}}$). For $K \in \mathcal{S}$, we denote by $K^{\uparrow}$ the set of statements of the form $[I : b \succ P]$ with $I \subseteq H$ and $P \subseteq O$ such that for all $i \in I$ and $c \in P$, $[i : b \succ c] \in K$. Conversely, for $\widehat{K} \in \widehat{\mathcal{S}}$, let $\widehat{K}^{\downarrow} = \{[i : b \succ c] \; : \; \exists [I : b \succ P] \in \widehat{K} \text{ s.t. } i \in I \text{ and } c \in P\}$ be the atomization of the factored statements $\widehat{K}$. Now assuming that a is the WCW, it is useful to distinguish different types of statements:

- positive statements, of the form $[I : a \succ P]$
- neutral statements, of the form $[I : b \succ P]$ with $a \notin P$
- negative statements, of the form $[I : b \succ P]$ with $a \in P$.

We note that in the case of basic statements, negative statements are "purely" negative since $P = \{a\}$.

Example 2. The full ranking of actions, on criterion 1 only, yields the following basic statements:

- $[1 : c \succ a]$ (negative statement),
- $[1 : c \succ e], [1 : c \succ d], [1 : c \succ b], [1 : c \succ f], [1 : e \succ d], [1 : e \succ b, [1 : e \succ f], [1 : d \succ b], [1 : d \succ f], [1 : b \succ f]$ (neutral statements),
- $[1 : a \succ e], [1 : a \succ d], [1 : a \succ b], [1 : a \succ f]$ (positive statements).

Regarding factored statements, the following examples can be given:

- $[1, 2 : e \succ d]$ is a neutral statement;
- $[1 : c \succ a, e]$ is a negative statement;
- $[1, 2 : a \succ d, e, f]$ is a positive statement.

The explanation shall also mention the weights in order to be complete. We assume throughout this paper that the values of weights can be shown to the audience. This is obvious in voting committee where the weights are public. This is also a reasonable assumption in a multi-criteria context when the weights are elicited, as the constructed weights are validated by the decision-maker and then become an important element of the explanation [6]. The corresponding language on the weights is simply composed of statements (called *importance statements*) of the form $[i : \alpha]$ with $i \in H$ and $\alpha \in [0, 1]$ meaning that the weight of criterion i is α. Let $\mathcal{W}$ (the set of normalized weights) be the set of sets $\{[i : w_i] \; : \; i \in H\}$ such that $w \in [0, 1]^H$ satisfies $\sum_{i \in H} w_i = 1$. For $W \in \mathcal{W}$ and $i \in H$, $W_i \in [0, 1]$ is the value of the weight on criterion i, that is that $[i : W_i] \in W$. A set $A \subseteq H$ is called a *winning coalition* if $\sum_{i \in A} W_i > \frac{1}{2}$.

2.3   Cost Function over the Explanations

An *explanation* is a pair composed of an element of $\widehat{\mathcal{S}}$ (note that $\mathcal{S} \subset \widehat{\mathcal{S}}$) and an element of $\mathcal{W}$. We seek for minimal explanations in the sense of some cost function. For simplicity, the cost of an element of $\widehat{\mathcal{S}}$ or $\mathcal{W}$ is assumed to be the sum of the cost of its statements. A difficult issue then arises: how should we define the cost of a statement?

Intuitively, the cost should capture the simplicity of the statement, the easiness for the user to understand it. Of course this cost must depend in the end of the basic pieces of information transmitted by the statement. The statements are of various complexity. For instance $[1, 2, 5, 7, 9 : a \succ b, c, g, h]$ looks more complex to grasp than $[1 : a \succ b]$, so that factored preference statements are basically more complex than basic preference statements.

Let us considered the case of preference statements. At this point we make the following assumptions:

- *neutrality*— the cost is insensitive to the identity of both criteria and options, i.e. $cost([I : b \succ P])$ depends only on $|I|$ and $|P|$ and is noted $C(|I|, |P|)$,
- *monotony*— the cost of a statement is monotonic w.r.t. criteria and to options, i.e. function C is non-decreasing in its two arguments. Neutrality implies that all basic statements have the same cost $C(1, 1)$.

Additionally to the previous properties, the cost may be sub-additive in the sense that $cost(I \cup I', P) \leq cost(I, P) + cost(I', P)$ and $cost(I, P \cup P') \leq cost(I, P) + cost(I, P')$, or super-additive if the converse inequalities hold. Finally, we assume the cost function can be computed in polynomial time.

3   Minimal Complete Explanations

Suppose now that the PI of choice problem is expressed in the basic language as a pair $\langle S, W \rangle \in \mathcal{S} \times \mathcal{W}$. Explaining why a is the Condorcet winner for $\langle S, W \rangle$ amounts to simplifying the PI (data-based approach [5]). We focus in this section on explanations in the language $\mathcal{S} \times \mathcal{W}$. The case of the other languages will be considered later in the paper.

A subset $\langle K, L \rangle$ of $\langle S, W \rangle$ is called a *complete explanation* if the decision remains unchanged regardless of how $\langle K, L \rangle$ is completed to form an element of $\mathcal{S} \times \mathcal{W}$. The completeness of the explanation is thus ensured. The pairs are equipped with the ordering $\langle K, L \rangle \sqsubseteq \langle K', L' \rangle$ if $K \subseteq K'$ and $L \subseteq L'$. More formally, we introduce the next definition.

Definition 2. *The set of* complete explanations *for language $\mathcal{S} \times \mathcal{W}$ is:*

$$Ex_{\mathcal{S},\mathcal{W}} := \{\langle K, L \rangle \sqsubseteq \langle S, W \rangle :$$
$$\forall K' \in \mathcal{S}(K) \; \forall L' \in \mathcal{W}(L) \quad WCW(K', L') = \{a\}\},$$

where $\mathcal{S}(K) = \{K' \in \mathcal{S} : K' \supseteq K\}$ and $\mathcal{W}(L) = \{L' \in \mathcal{W} : L' \supseteq L\}$.

Example 3. The explanation $K_1 = [1, 2 : a \succ d, e, f], [1, 3 : a \succ b], [2, 3 : a \succ c]$ is not complete, since it does not provide enough evidence that a is preferred over c. Indeed, $H_{K_1}(a, c) < 0$ (since $0.42 - 0.58 = -0.16$). On the other hand, $[1 : a \succ e, d, b, f], [2 : a \succ f, e, d, c], [3 : a \succ b, c, d], [4 : a \succ c, e], [5 : a \succ c]$ is complete but certainly not minimal, since (for instance) exactly the same explanation without the last statement is also a complete explanation whose cost is certainly lower (by monotonicity of the cost function). Now if the cost function is sub-additive, then a minimal explanation cannot contain (for instance) both $[1, 2 : a \succ d, e]$ and $[1, 2 : a \succ f]$. This is so because then it would be possible to factor these statements as $[1, 2 : a \succ d, e, f]$, all other things being equal, so as to obtain a new explanation with a lower cost.

In the rest of the paper, complete explanations will be called simply explanations when there is no possible confusion. One has $\langle S, W \rangle \in Ex_{S,W}$ and $\langle \emptyset, \emptyset \rangle \notin Ex_{S,W}$. As shown below, adding more information to a complete explanation also yields a complete explanation.

Lemma 1. *If $\langle K, L \rangle \in Ex_{S,W}$ then $\langle K', L' \rangle \in Ex_{S,W}$ for all K', L' with $K \subseteq K' \subseteq S$ and $L \subseteq L' \subseteq W$.*

Proof : Clear since $\mathcal{S}(K) \supseteq \mathcal{S}(K')$ when $K \subseteq K'$, and $\mathcal{W}(L) \supseteq \mathcal{W}(L')$ when $L \subseteq L'$. ∎

We will assume in the rest of the paper that there is no simplification regarding the preferential information W. Indeed the gain of displaying less values of the weights is much less significant than the gain concerning S. This comes from the fact that $|W| = m$ whereas $|S| = \frac{1}{2} m \, p \, (p - 1)$, where $m = |H|$ and $p = |O|$. Only the information about the basic statements $S \in \mathcal{S}$ is simplified. We are thus interested in the elements of $Ex_{S,W}$ of the form $\langle K, W \rangle$. Hence we introduce the notation $Ex_S = \{K \in \mathcal{S} \; : \; \langle K, W \rangle \in Ex_{S,W}\}$.

4   Simple Language for S

We consider in this section explanations with the basic languages $\mathcal{S}$ and $\mathcal{W}$. In this section, the PI is expressed as $\langle S, W \rangle$. The aim of this section is to characterize and construct minimal elements of Ex_S w.r.t. the cost.

We set $H_K(a, b) := \sum_{i \, : \, [i : a \succ b] \in K} W_i - \sum_{i \, : \, [i : a \succ b] \notin K} W_i$ for $K \subseteq S$ and $b \in O^\star$. This means that $K \subseteq S$ is completed only with negative preference statements (in other words, what is not explicitly provided in the explanation is assumed to be negative).

Lemma 2. $Ex_S = \{K \subseteq S \; : \; \forall b \in O^\star \quad H_K(a, b) > 0\}$.

Proof : We have $\text{WCW}(K', W) = \{a\} \; \forall K' \in \mathcal{S}(K)$ iff $\text{WCW}(K', W) = \{a\}$ for $K' = K \cup \{[i : b \succ a] \; : \; b \in O^\star \text{ and } [i : a \succ b], [i : b \succ a] \notin K\}$ iff $H_K(a, b) > 0$ $\forall b \in O^\star$. ∎

A consequence of this result is that neutral statements can simply be ignored since they do not affect the expression $H_K(a, b)$. The next lemma shows furthermore that the minimal explanations are free of negative statements.

Lemma 3. *Let $K \in Ex_S$ minimal w.r.t. the cost. Then K does not contain any negative or neutral preference statement.*

Proof : $K \in Ex_S$ cannot minimize the cost if $[i : b \succ a] \in K$ since then $H_{K'}(a, b) = H_K(a, b)$ and thus $K' \in Ex_S$, with $K' = K \setminus \{[i : b \succ a]\}$. It is the same if $[i : b \succ c] \in K$ with $b, c \neq a$. ■

Then we prove that we can replace a positive basic statement appearing in a complete explanation by another one, while having still a complete explanation, if the weight of the criterion involved in the first statement is not larger than that involved in the second one.

Lemma 4. *Let $K \in Ex_S$, $[i : a \succ b] \in K$ and $[j : a \succ b] \in S \setminus K$ with $W_j \geq W_i$. Then $(K \setminus \{[i : a \succ b]\}) \cup \{[j : a \succ b]\} \in Ex_S$.*

Proof : Let $K' = (K \setminus \{[i : a \succ b]\}) \cup \{[j : a \succ b]\}$. We have $H_{K'}(a, b) = H_K(a, b) + 2(W_j - W_i) > 0$. Hence $K' \in Ex_S$. ■

We define $\Delta_i^S(a, b) = +1$ if $[i : a \succ b] \in S$, and $\Delta_i^S(a, b) = -1$ if $[i : b \succ a] \in S$. For each option $b \in O^\star$, we sort the criteria in H by a permutation π_b on H such that $W_{\pi_b(1)} \Delta_{\pi_b(1)}^S(a, b) \geq \cdots \geq W_{\pi_b(m)} \Delta_{\pi_b(m)}^S(a, b)$.

Proposition 1. *For each $b \in O^\star$, let p_b the smallest integer such that $H_{K_{p_b}^b}(a, b) > 0$, where $K_{p_b}^b = \{[\pi_b(1) : a \succ b], [\pi_b(2) : a \succ b], \ldots, [\pi_b(p_b) : a \succ b]\}$. Then $\{[\pi_b(j) : a \succ b] : b \in O^\star \text{ and } j \in \{1, \ldots, p_b\}\}$ is a minimal element of Ex_S w.r.t. the cost.*

Proof (Sketch): Let $Ex_S(b) = \{K \subseteq S_b : H_K(a, b) > 0\}$, where S_b is the set of statements of S involving option b. The existence of p_b follows from the fact that a is a WCW. Now let $j \in \{1, \ldots, p_b - 1\}$. From the definition of p_b, $K_{p_b - 1}^b \notin Ex_S(b)$. This, together with $W_{\pi_b(j)} \geq W_{\pi_b(p_b)}$ and Lemma 4, implies that $K_{p_b}^b \setminus \{[\pi_b(j) : a \succ b]\} \notin Ex_S(b)$. Hence $K_{p_b}^b$ is minimal in $Ex_S(b)$ in the sense of $\subseteq$. It is also apparent from Lemma 4 that there is no element of $Ex_S(b)$ with a strictly lower cardinality and thus lower cost (since, from Section 2.3, the cost of a set of basic statements is proportional to its cardinality). Finally, $\cup_{b \in O^\star} K_{p_b}$ minimizes the cost in Ex_S since the conditions on each option $b \in O^\star$ are independent. ■

This proposition provides a polynomial computation of a minimal element of Ex_S. This is obtained for instance by the following greedy Algorithm 1. The complexity of this algorithm is $O(m \cdot p \cdot \log(p))$ (where $m = |H|$ and $p = |O|$).

Function Algo(W, Δ) :
    $K = \emptyset$;
    For each $b \in O^\star$ **do**
        Determine a ranking π_b of the criteria according to $W_j \Delta_j^S(a, b)$ such
        that $W_{\pi_b(1)} \Delta_{\pi_b(1)}^S(a, b) \geq \cdots \geq W_{\pi_b(m)} \Delta_{\pi_b(m)}^S(a, b)$;
        $K_b = \{[\pi_b(1) : a > b]\}; k = 1$;
        While $(H_{K_b}(a, b) \leq 0)$ **do**
            $k = k + 1; K_b = K_b \cup \{[\pi_b(k) : a > b]\}$;
        done
        $K = K \cup K_b$;
    end For
    return K;
End

Algorithm 1. Algorithm for the determination of a minimal element of $Ex_{\mathcal{S}}$. The outcome is K.

We illustrate this on our example.

Example 4. Consider the iteration regarding option b. The ranking of criteria for this option is $1/3/4/5/2$. During this iteration, the statements $[1 : a \succ b], [3 : a \succ b]$ are added to the explanation. In the end the explanation produced by Algorithm 1 is $[1 : a \succ b], [3 : a \succ b], [2 : a \succ c], [3 : a \succ c], [4 : a \succ c], [1 : a \succ d], [2 : a \succ d], [1 : a \succ e], [2 : a \succ e], [1 : a \succ f], [2 : a \succ f]$. Note that criterion 5 is never involved in the explanation.

5   Factored Language for S

The language used in the previous section is simple but not very intuitive. As illustrated in the introduction, a natural extension is to allow more compact explanations by means of factored statements. We thus consider in this section explanations with the factored language $\widehat{\mathcal{S}}$ and the basic language $\mathcal{W}$. As in previous section, all weight statements in $W \in \mathcal{W}$ are kept. The explanations for $\widehat{\mathcal{S}}$ are:

$$Ex_{\widehat{\mathcal{S}}} = \left\{ \widehat{K} \subseteq S^\uparrow \; : \; \forall K \in \mathcal{S}(\widehat{K}^\downarrow) \quad \text{WCW}(K, W) = \{a\} \right\}.$$

Similarly to what was proved for basic statements, it is simple to show that minimal explanation must only contain positive statements.

Lemma 5. *Let $\widehat{K} \in Ex_{\widehat{\mathcal{S}}}$ minimal w.r.t. the cost. Then $\widehat{K}$ only contains positive preference statements.*

Proof : Similar to the proof of Lemma 3. ∎

A practical consequence of this result is that it is sufficient to represent the PI as a binary matrix, for a, where an entry 1 at coordinates (i, j) represents the

fact that the option i is less preferred than a on criteria j. Doing so, we do not encode the preferential information expressed by neutral statements.

This representation is attractive because factored statements visually correspond to (combinatorial) rectangles. Informally, looking for an explanation amounts to find a "cheap" way to "sufficiently" cover the 1's in this matrix. However, an interesting thing to notice is that a minimal explanation with factored statements does not imply that factored statements are non overlapping. To put it differently, it may be the case that some preferential information is repeated in the explanations. Consider the following example:

Example 5. There are 5 criteria of equal weight and 6 options, and a is the weighted Condorcet winner. As for the cost of statements, it is constant whatever the statement.

	1	2	3	4	5
	0.2	0.2	0.2	0.2	0.2
b	1	1	0	0	1
c	1	1	0	1	0
d	1	1	1	0	0
e	0	1	1	0	1
f	0	1	1	1	0

There are several minimal explanations involving 4 statements, but all of them result in a covering in the matrix, like for instance $[1, 2 : a \succ b, c, d], [2, 3 : a \succ d, e, f], [4 : a \succ c, f][5 : a \succ b, e]$, where the preferential information that $a \succ_2 d$ is expressed twice (in the first and second statement).

The previous section concluded on a simple algorithm to compute minimal explanations with basic statements. Unfortunately, we will see that the additional expressive power provided by the factored statements comes at a price when we want to compute minimal explanations.

Proposition 2 (Min. explanations with factored statements). *Deciding if (using factored statements $S^\uparrow$) there exists an explanation of cost at most k is NP-complete. This holds even if criteria are unweighted and if the cost of any statement is a constant.*

Proof (Sketch): Membership is direct since computing the cost of an explanation can be done in polynomial time. We show hardness by reduction from the BICLIQUE EDGE COVER (BEC), known to be NP-complete (problem [GT18] in [7]). In BEC, we are given a finite bipartite graph $G = (X, Y, E)$ and positive integer k'. A biclique is a complete bipartite subgraph of G, *i.e.*, a subgraph induced by a subset of vertices such that any vertex is connected to a vertex of the other part. The question is whether there exists a collection of bicliques covering edges of G of size at most k'.

Let $I = (X, Y, E)$ be an instance of BEC. From I, we build an instance I' of the explanation problem as follows. The set O of actions contains $O_1 = \{o_1, \ldots, o_n\}$ corresponding to the elements in X, and a set O_2 of dummy actions consisting

of $n+3$ actions $\{o'_1, \ldots, o'_{n+3}\}$. The set H of criteria contains $H_1 = \{h_1, \ldots, h_n\}$ corresponding to the elements in Y, and a set H_2 of dummy criteria consisting of $n+3$ criteria $\{h'_1, \ldots, h'_{n+3}\}$. First, for each $(x_i, y_j) \in E$, we build a statement $[h_i : a \succ o_j]$. Let S_{O_1, H_1} be this set of statements. Observe that a factored statement $[I : a \succ P]$ with $I \subseteq H_1$ and $P \subseteq O_1$ correspond to a biclique in I. But a may not be a Condorcet winner. Thus for each action $o \in O_1$, we add $(n+2) - |\{[h_i : a \succ o] \in O_1\}|$ statement(s) $[h'_j : a \succ o]$. Let S_{O_1, H_2} be this set of statements. Note that at this point, a is preferred to any other $o \in O_1$ by $n+2$ criteria. Next $\forall (h'_i, o'_j) \in (H_2 \times O_2)$ such that $i \neq j$ we add the following statement: $[h'_i : a \succ o'_j]$. There are $n+2$ such statements, hence a is preferred to any other $o \in O_2$ by a majority of exactly $n+2$ criteria. Let S_{O_2, H_2} be this set of statements. We claim that I admits a biclique vertex partition of at most $k - (n+3)$ subsets iff I' admits an explanation $\widehat{K}_*$ of cost at most k using factored statements. Take ($\Leftarrow$). By construction, all the basic statements must be "covered", i.e. $\widehat{K}_*^{\downarrow} = S_{O_1, H_1} \cup S_{O_1, H_2} \cup S_{O_2, H_2}$. We denote by $cov(.)$ the cost of covering a set of basic statements of $S_{O,H}$ (this is just the number of factored statements used, as the cost of statements is constant). Furthermore, as there are no statements using actions from O_2 and criteria from H_1, no factored statement can cover at the same time statements from S_{O_1, H_1} and S_{O_2, H_2}. Hence $cost(\widehat{K}_*) = cov(S_{O_1, H_1} \cup S') + cov(S_{O_2, H_2} \cup S'')$, such that $S' \cup S'' = S_{O_1, H_2}$.

But now observe that $cov(S_{O_2, H_2}) = cov(S_{O_2, H_2} \cup S_{O_1, H_2}) = n+3$, so $cost(\widehat{K}_*)$ boils down to $n+3 + cov(S_{O_1, H_1} \cup S')$. By monotony wrt. criteria, $cov(S_{O_1, H_1} \cup S')$ is minimized when $S' = \emptyset$, and this leads to the fact $cov(S_{O_1, H_1}) \leq k - (n+3)$. The ($\Rightarrow$) direction is easy. ∎

The previous result essentially shows that when the cost function implies to minimize the number of factored statements, no efficient algorithm can determine minimal explanations (unless P=NP). But there may be specific class(es) of cost functions for which the problem may turn out to be easy. As shown in the next lemma, when the cost function is super-additive, then it is sufficient to look for basic statements.

Lemma 6. *If the cost function is super-additive, then $\min_{\widehat{K} \in Ex_{\widehat{S}}} cost(\widehat{K}) = \min_{K \in Ex_S} cost(K)$.*

Proof : Let $\widehat{K} \in Ex_{\widehat{S}}$. We know that $\widehat{K}^{\downarrow} \in Ex_S$. By super-additivity, $cost(\widehat{K}) = \sum_{[I:b \succ P] \in \widehat{K}} cost([I : b \succ P]) \geq \sum_{[I:b \succ P] \in \widehat{K}} \sum_{i \in I, c \in P} cost([i : b \succ c]) \geq \sum_{[i:b \succ c] \in \widehat{K}^{\downarrow}} cost([i : b \succ c]) = cost(\widehat{K}^{\downarrow})$. ∎

Yet, the cost is expected to be sub-additive. Relations (1) and (2) below give examples of sub-additive cost functions. In this case, factored statements are less costly (e.g. the cost of $[\{1, 2\} : a \succ b]$ should not be larger than the cost of $[1 : a \succ b], [2 : a \succ b]$) and factored explanations become very relevant.

When the cost function is sub-additive, an intuitive idea could be to restrict our attention to statements which exhibit winning coalitions. For that purpose, let us assign to any subset $P \subseteq O^\star$ defended by a winning coalition the cost

of using such statement. A practical way to do this is to build $T : 2^{O^\star} \to 2^H$ such that for all subsets $P \subseteq O^\star$, $T(P)$ is the largest set of criteria for which $[T(P) : a \succ P] \in S^\uparrow$. We have $T(P) = \cap_{b \in P} T(\{b\})$, where $T(\{b\}) := \{i \in H : [i : a \succ b] \in S\}$. Then subsets P of increasing cardinality are considered (but those supported by non-winning coalitions are discarded). The cost $C(\alpha, |P|)$ is finally assigned, where α is the size of the smallest winning coalition contained in $T(P)$. Then, the problem can be turned into a weighted set packing, for which the direct ILP formulation would certainly be sufficient in practice for reasonable values of $|O|$ and $|H|$.

Example 6. On our running example, the different potential factors would be $T(\{b\}) = \{1, 3\}$ with $C(2, 1)$, $T(\{c\}) = \{2, 3, 4, 5\}$ with $C(4, 1)$, $T(\{d\}) = \{1, 2, 3\}$ with $C(3, 1)$, $T(\{e\}) = \{1, 2, 4\}$ with $C(3, 1)$, $T(\{f\}) = \{1, 2\}$ with $C(2, 1)$, $T(\{b, d\}) = \{1, 3\}$ with $C(2, 2)$, etc. Depending on the cost function, two possible explanations remain: $\widehat{K}_1 = \{[1, 3 : a \succ b], [2, 3, 4, 5 : a \succ c], [1, 2 : a \succ d, e, f]\}$ for a cost of $C(2, 1) + C(4, 1) + C(2, 3)$, and $\widehat{K}_2 = \{[1, 3 : a \succ b, d], [2, 3, 4, 5 : a \succ c], [1, 2 : a \succ e, f]\}$ for a cost of $C(2, 2) + C(4, 1) + C(2, 2)$. The cost function

$$C(i, j) = i^\alpha j^\beta \tag{1}$$

(which is sub-additive when $\alpha \leq 1$ and $\beta \leq 1$) would select $\widehat{K}_1$. Note that criteria 4 or 5 will be dropped from the statement $[T(\{c\}) : a \succ c]$.

Now, considering only factored statements with winning coalitions may certainly prevent from reaching optimal factored explanations, as we illustrate below.

Example 7. We have 4 criteria and 3 options. Assume that a is preferred to b on criteria $1, 2$, and 3; that a is preferred to c on criteria $1, 2$, and 4 and that any coalition of at least 3 criteria is winning. The previous approach based on T gives $\widehat{K}_1 = \{[1, 2, 3 : a \succ b], [1, 2, 4 : a \succ c]\}$, with $cost(\widehat{K}_1) = 2\, C(3, 1)$. Algorithm 1 gives $\widehat{K}_2 = (\widehat{K}_1)^\downarrow$ with $cost(\widehat{K}_2) = 6\, C(1, 1)$. Another option is to consider $\widehat{K}_3 = \{[1, 2 : a \succ b, c], [3 : a \succ b][4 : a \succ c]\}$, with $cost(\widehat{K}_3) = C(2, 2) + 2\, C(1, 1)$. Let us consider the following cost function[1]

$$C(i, j) = i \log(j + 1). \tag{2}$$

Function C is sub-additive, since $C(i + i', j) = C(i, j) + C(i', j)$ and, from relation $j + j' + 1 \leq (j + 1)(j' + 1)$, we obtain $C(i, j + j') \leq C(i, j) + C(i, j')$. Then we have $cost(\widehat{K}_3) < cost(\widehat{K}_1) = cost(\widehat{K}_2)$ so that the explanation with the smallest cost is $\widehat{K}_3$.

Enforcing complete explanations implies a relatively large number of terms in the explanation. However, in most cases, factored statements allow to obtain small explanations. For instance, when all criteria have the same weight, the minimal elements of Ex_S contain exactly $(p - 1)\, n$ basic statements (where $p = |O|$,

---

[1] Capturing that factoring over the criteria is more difficult to handle than factoring over the options.

$m = |H|$ and $m = 2n - 1$ if m is odd, and $m = 2n - 2$ if m is even. Indeed, one needs $p - 1$ terms to explain that a is globally preferred over b, for all $b \in O^\star$, and the minimal elements of $Ex_{\widehat{S}}$ contain at most $p - 1$ factored statements (factoring with winning coalitions for each $b \in O^\star$).

A current matter of investigation is to determine the class of cost functions for which the minimal explanation is not given either by trivial atomization or by factoring with winning coalitions only, thus requiring dedicated algorithms.

6   Related Work and Conclusion

The problem of producing explanations for complex decisions is a long-standing issue in Artificial Intelligence in general. To start with, it is sometimes necessary to (naturally) explain that no satisfying option can be found because the problem is over-constrained [8,9]. But of course it is also important to justify why an option is selected among many other competing options, as is typically the case in recommendations. Explanations based on the data seek to focus on a small subpart of the data, sufficient to either convince or indeed prove the claim to the user. Depending on the underlying decision model, this can turn out to be very challenging.

In this paper we investigate the problem of providing minimal and complete explanations for decisions based on a weighted majority principle, when a Condorcet winner exists. A first contribution of this paper is to set up the framework allowing to analyze notions of minimal explanations, introducing in particular different languages to express the preferential information. We then characterize minimal explanations, and study their computational properties. Essentially, we see that producing minimal explanations is easy with basic statements but may be challenging with more expressive languages.

Much work in argumentation set up theoretical systems upon which various types of reasoning can be performed, in particular argument-based decision-making has been advocated in [10]. The perspective taken in this paper is different in at least two respects: (i) the decision model is not argumentative in itself, the purpose being instead to generate arguments explaining a multiattribute decision model (weighted majority) issued from decision-theory; and (ii) the arguments we produce are complete (so, really *proving* the claim), whereas in argumentation the defeasible nature of the evidence put forward is a core assumption [11]. Regarding (ii), our focus on complete arguments has been justified in the introduction. Regarding (i), we should emphasize that we make no claim on the relative merits of argument-based vs. decision-theoretic models. But in many organizations, these decision models are currently in use, and although it may be difficult to change the habits of decision-makers for a fully different approach, adding explanatory features on top on their favorite model can certainly bring much added-value. This approach is not completely new, but previous proposals are mainly heuristic and seek to generate natural arguments [1] that are persuasive in practice. An exception is the recent proposal of [6] which provides solid theoretical foundations to produce explanations for a range

of decision-theoretic weight-based models, but differs in (ii) since explanations are based on (defeasible) argument schemes. Our focus on complete explanations is a further motivation to build on solid theoretical grounds (even though weaker incomplete arguments may prove more persuasive in practice).

Recently, the field of computational social choice has emerged at the interface of AI and social choice, the study of various computational of various voting systems being one of the main topic in this field. There are connections to our work (and indeed one of the motivating example is a voting committee): for instance, exhibiting the smallest subsets of votes such that a candidate is a *necessary winner* [12] may be interpret as a minimal (complete) explanation that this candidate indeed wins. However, the typical setting of voting (*e.g.* guaranteeing the anonymity of voters) would not necessarily allow such explanations to be produced, as it implies to identify voters (to assign weights). An interesting avenue for future research would be to investigate what type of explanations would be acceptable in this context, perhaps balancing the requirements of privacy and the need to support the result. We believe our approach could be relevant. Indeed, two things are noteworthy: first, the proposed approach already preserves some privacy, since typically only parts of the ballots need to be exhibited. Secondly, in many cases it would not be necessary to exactly identify voters, at least when their weights are sufficiently close. Take again our running example: to explain that a beats b we may well say "the most important voter 1 is for a, and among 2 and 3 only one defends b".

We conclude by citing some possible extensions of this work. The first is to improve further the language used for explanations. The limitations of factored statements is clear when the following example is considered:

Example 8. In the following example with 6 alternatives and 5 criteria (with the same weight), the factored statements present in any minimal explanation contain at least 3 criteria or alternatives (for instance, $[1, 2, 3 : a \succ e, f], [3, 4, 5 : a \succ b, c], [1, 2, 4 : a \succ d]$)

1	2	3	4	5
0.2	0.2	0.2	0.2	0.2
b	c	d	e	f
a	a	a	a	a
c	d	e	f	b
d	e	f	b	c
e	f	b	c	d
f	b	c	d	e

However, an intuitive explanation that comes directly to mind is as follows: "a is only beaten by a different option on each criteria".

To take a step in the direction of such more natural explanations, the use of "except" statements allowing to assert that an option is preferred over any other option *except* the ones explicitly cited should be taken into account. (In fact, the informal explanation of our example makes also use of such a statement, since

it essentially says that a is better than c on all criteria except 1). In that case, minimal explanations may cover larger sets of basic statements than strictly necessary (since including more elements of the PI may allow to make use of an except statement). Another extension would be to relax the assumption of neutrality of the cost function, to account for situations where some information is exogenously provided regarding criteria to be used preferably in the explanation (this may be based on the profile of the decision-maker, which may be more sensible to certain types of criteria).

References

1. Carenini, G., Moore, J.: Generating and evaluating evaluative arguments. Artificial Intelligence 170, 925–952 (2006)
2. Klein, D.: Decision analytic intelligent systems: automated explanation and knowledge acquisition. Lawrence Erlbaum Associates, Mahwah (1994)
3. Buchanan, B.G., Shortliffe, E.H.: Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Boston (1984)
4. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: MoviExplain: a recommender system with explanations. In: Proceedings of the Third ACM Conference on Recommender Systems (RecSys 2009), pp. 317–320. ACM, New York (2009)
5. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2000), pp. 241–250. ACM, New York (2000)
6. Labreuche, C.: A general framework for explaining the results of a multi-attribute preference model. Artificial Intelligence 175, 1410–1448 (2011)
7. Garey, M., Johnson, D.: Computers and intractability. A guide to the theory of NP-completeness. Freeman, New York (1979)
8. Junker, U.: QUICKXPLAIN: Preferred explanations and relaxations for over-constrained problems. In: McGuinness, D.L., Ferguson, G. (eds.) Proceedings of the Nineteenth AAAI Conference on Artificial Intelligence (AAAI 2004), pp. 167–172. AAAI Press, Menlo Park (2004)
9. O'Sullivan, B., Papadopoulos, A., Faltings, B., Pu, P.: Representative explanations for over-constrained problems. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI 2007), pp. 323–328. AAAI Press, Menlo Park (2007)
10. Amgoud, L., Prade, H.: Using arguments for making and explaining decisions. Artificial Intelligence 173, 413–436 (2009)
11. Loui, R.P.: Process and policy: Resource-bounded nondemonstrative reasoning. Computational Intelligence 14, 1–38 (1998)
12. Konczak, K., Lang, J.: Voting procedures with incomplete preferences. In: Brafman, R., Junker, U. (eds.) Proceedings of the IJCAI 2005 Workshop on Advances in Preference Handling, pp. 124–129 (2005)

C.3 Selection of articles related to Chapter 5

- Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane, Simon Parsons. A dialogue game for recommendation with adaptive preference models. In proceeding of the 14th International Conference on Autonomous Agents and Multiagent systems. Istanbul, Turkey. pp.959-967. 2015.

- Wassila Ouerdane, Yannis Dimopoulos, Konstantinos Liapis, Pavlos Moraitis. Towards automating Decision Aiding through Argumentation. Journal of Multicriteria Decision Analysis, Volume 18, pp 289-309, 2011.

- Wassila Ouerdane, Nicolas Maudet, Alexis Tsoukiàs. Argument Schemes and Critical Questions for Decision Aiding Process. Proceedings of the 2nd international conference on Computational Models of Argument (COMMA2008), Toulouse, France. pp. 285-296, 2008

A Dialogue Game for Recommendation with Adaptive Preference Models

C. Labreuche
Thales Research & Technology
91767 Palaiseau Cedex
France
christophe.labreuche@thalesgroup.com

N. Maudet
Sorbonne Universités, UPMC
Univ Paris 06
CNRS, UMR 7606, LIP6
F-75005, Paris, France
nicolas.maudet@lip6.fr

W. Ouerdane
LGI, CentraleSupélec
Chatenay Malabry
France
wassila.ouerdane@ecp.fr

S. Parsons
Department of Computer
Science
University of Liverpool
UK
s.d.parsons@liverpool.ac.uk

ABSTRACT

To provide convincing recommendations, which can be fully understood and accepted by a decision-maker, a decision-aider must often engage in an interaction and take the decision maker's responses into account. This feedback can lead to revising the model used to represent the preferences of the decision-maker. Our objective in this paper is to equip an artificial decision-aider with this adaptive behavior. To do that, we build on decision theory to propose a principled way to select decision models.

Our approach is axiomatic in that it does not only work for a predefined subset of methods—we instead provide the properties that make models compatible with our proposal. Finally, the interaction model is complex since it can involve the exchange of different types of preferential information, as well as others locutions such as justifications. We manage it through a dialogue game, and prove that it satisfies desired properties, in particular termination, and efficiency

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Coherence and coordination, intelligent agents, Multiagent systems*

Keywords

Recommender system; Communication protocol; Argumentation

1. INTRODUCTION

In a decision aiding context, there are at least two distinct actors: a *decision maker* (DM), and an analyst, that we call a *decision aider* (DA). These play very different roles [24]. The DM explains the decision problem to the DA, has some

preferences on the decision options and is at the end responsible for the decision and its justifcation. The DA helps him in this task by bringing some methodology and rationality. The DA analyses the consistency of the information provided by the DM, proposes some recommendation on the basis of such information and constructs the corresponding justifications. A key ingredient of the decision process is how interaction takes place. In particular, the DA should be able to adapt itself to the responses of the DM. In fact, the DM's preferences are often incomplete, or at least not fixed at the beginning of the process. Only when confronted with the recommendation can the DM react and give feedback. The competence of a human DA is precisely to integrate this new information, to revise her representation of the profile of the DM, so as to produce a finely adapted recommendation which can be understood and accepted.

This raises a challenging issue when the DA is an artificial agent, since it must have precisely this ability to adapt itself to the responses of the DM. Take for instance recommender systems used in commercial websites: the role of the DA is to recommend items that the DM is likely to buy (travel, books, etc.). Often the product space is extremely large, and the role of the DA is to help to navigate in this catalogue. According to [14], "user feedback is a vital component of most recommenders". In recommender systems, this feedback of the DM can take various forms: value-based feedback which asserts a value on a given attribute ("I want three gears on my bike"), preference-based feedback which singles out a favorite item so as to get more of the same type in the next cycle of recommendation ("This is the bike I prefer, can you show more like this?"), or critique-based feedback, which can be seen as a mixture of the two since the DM picks a preferred item but at the same expresses how it could be improved ("This type of bike, but in a different color."). Many recommender systems do not explicitly construct a preference model, and thus have no memory of user feedback. The system can then recommend an option which the user crticised a few iterations before. To take proper account of user feedback in timely and consistent manner, some authors argue to maintain the user's preference model [5, 19, 25]. Model-based recommendation systems are then based on a unique model (e.g. additive utility) and rely upon

the assumption that all potential users can be represented by this model [4, 25]. However, in the case of multi-criteria recommendation, there is a wide variety of possible preference models, and assuming a fixed model may prove to be too restrictive. Suppose for instance that the DA starts with a majority model, but later realizes that the user shall be represented by quantitative utilities and thus switches to additive utility model.

In this paper we consider a simple recommendation scenario where a set of available options is known at the start. To remedy a previous flaw, here we propose to *allow an artificial DA to use a variety of decision models* (able to encompass most of decision situations) to build its recommendation (as opposed to adjusting the parameters of a single model). This raises some obvious questions: *(i)* if the DA can choose among several models, is there a principled way to do so? *(ii)* would such a method be dependent of the models considered? And, finally *(iii)* how, in practice, should such an interaction be regulated?

We borrow from decision theory and Multiple Criteria Decision Analysis to answer the first point in the positive. Regarding *(ii)*, we advocate a generic method to account for this adaptative behavior. Indeed, instead of focusing on a given collection of models, we adopt an axiomatic approach, and thus characterize which models can be handled in the way we propose. As for *(iii)*, the actual procedure we put forward takes the form of a dialogue game between the DM and the DA, and is inspired by recent work in dialectical models of interaction.

2. A MOTIVATING EXAMPLE

We give a basic dialogue example between a DA and a DM for a given decision problem. It shows the asymmetry between the two agents: on the one hand the different feedback expressed by the DM (asking explanations, suggesting new propositions, etc.), and on the other hand how the DA adapts its responses (providing justifications, changing or correcting the recommendation, etc.).

EXAMPLE 1. *Suppose that a DM specifies that he has to rank four options $\{a, b, c, d\}$ (say, bikes to be deployed for sharing in big city). Each bike is evaluated on the set $\{c_1, c_2, c_3, c_4, c_5\}$ of criteria (say, price, weight, esthetic, gears, dimension). The comparison of the options over each criteria (where $x \succ_{c_i} y$ means that option x is strictly preferred to y on criterion c_i) is as follows:*

$$c_1: d \succ_{c_1} a \succ_{c_1} c \succ_{c_1} b;$$
$$c_2: d \succ_{c_2} a \succ_{c_2} b \succ_{c_2} c;$$
$$c_3: b \succ_{c_3} c \succ_{c_3} a \succ_{c_3} d;$$
$$c_4: c \succ_{c_4} b \succ_{c_4} a \succ_{c_4} d;$$
$$c_5: b \succ_{c_5} a \succ_{c_5} c \succ_{c_5} d.$$

On the basis of this information, the DA computes a recommendation and submits it to the DM for a discussion. Such a discussion unfolds as follows:

(1) DA: I recommend that $b \succ a \succ c \succ d$.
(2) DM: Why $b \succ a$?
(3) DA: b is better on a majority of criteria (c_3, c_4, c_5).
(4) DM: I see, but still I would prefer a to b
(5) DA: Why?
(6) DM: Because a is better on the price and weight (c_1, c_2), these are very important criteria.
(7) DA: Fine. I still recommend b over c.
(8) DM: Yes, I agree.

(9) DA: Now, would d be preferred to a?
(10) DM: No, this is not the case.
(11) DA: Why?
(12) DM: Because d is bad on esthetic, gears and dimension (c_3, c_4, c_5).
(13) DA: Ok, so c is better than d.
(14) DM: I agree.
(15) DA: Do you agree also that a is preferred to b?
(16) DM: Of course.
(17) DA: Ok. My recommendation is thus: $a \succ b \succ c \succ d$.

3. BASIC DEFINITIONS

We consider a finite set O of options, a finite set H of criteria. The recommendations of the DA are based on a *decision model*, which provides a total order of the elements of O on the basis of their evaluations on the criteria. There are many different decision models in the literature. Each model corresponds to different rationality assumptions on the DM. Since neither DA nor DM know in advance what model best represents the DM, one cannot use a single predefined decision model. Rather we use a family Π of decision models that encompasses most commonly encountered DM profiles. In order to support our running example, we consider four decision models (described formally below), but our approach is not restricted to these models.

EXAMPLE 2. *In the rest of the paper, for illustration, we will consider the following family Π of models: Simple Majority model (noted π_{SM}), Simple Weighted Majority model (π_{SWM}), Mean model (π_M) and Weighted Sum model (π_{WS}).*

3.1 Description of the preference information

In order to make a decision between several options, the DM needs to provide information about the evaluation of an option $x \in O$, and about the relative strength of criteria. We will make use of two evaluation scales:

- an evaluation scale for the options on the criteria $\mathcal{S}_O$, e.g. $\mathcal{S}_O = \{$ good, average, bad$\}$;

- an evaluation scale for the importance of criteria $\mathcal{S}_H$, e.g. $\mathcal{S}_H = \{$ strong, average, weak$\}$.

The DM expresses some *preference information* (PI) which is related to the comparison of the options on the criteria, or the importance of criteria. This PI allows to construct a preference relation among the options, thanks to the use of a model in Π. The PI is expressed by means of different types of statements:

DEF. 1. *An* evaluation statement *is of the form $[c : x = \alpha]$ where $x \in O$, $c \in H$ and $\alpha \in \mathcal{S}_O$, meaning that the assessment of option x on criterion c is equal to α.*

DEF. 2. *A* preference statement *is of the form $[x \succ_c y]$ where $x, y \in O$ and $c \in H$, meaning that x is preferred to y on criterion c.*

DEF. 3. *A* weight statement *is of the form $[c = \alpha]$ where $c \in H$ and $\alpha \in \mathcal{S}_H$, meaning that the importance of criterion c is equal to α.*

EXAMPLE 3. *(Ex. 1, cont.) We have many preference statements of the form $[d \succ_{c_1} a]$. In Turn 12, the DM uses an evaluation statement: $[c_4 : d = bad]$, while in Turn 6, the DM uses a weight statement: $[c_1 = strong]$, $[c_2 = strong]$.*

In order to make inferences from PI, this latter shall be consistent. This concept is now defined.

DEF. 4. *The previous statements are called* PI statements. *A subset P of PI statements is said to be* consistent *if there is no two evaluation statements $[c : x = \alpha], [c : x = \alpha']$ with $\alpha \neq \alpha'$, there is no cycle of $\succ_c$ for preference statements, and there is no two weight statements $[c = \alpha], [c = \alpha']$ with $\alpha \neq \alpha'$.*

Clearly, the use of some type of statements says something about the underlying preference model. Let $\mathcal{P}(\pi)$, with $\pi \in \Pi$, denote the set of such statements that can be used for constructing model π (see Ex. 4 below), and $\mathcal{P} = \cup_{\pi \in \Pi} \mathcal{P}(\pi)$. Thus we have:

DEF. 5. *The Preference Information (PI) is any subset of $\mathcal{P}$. The Preference Information (PI) for a decision model $\pi \in \Pi$ is any subset $P \subseteq \mathcal{P}(\pi)$.*

The value of $\mathcal{P}(\pi)$ for the different models is now shown in the four models.

EXAMPLE 4. *(Ex. 2 Cont.)*

- *the model π_{SM} relies only on the preference statements: $\mathcal{P}(\pi_{SM}) = \{[a \succ_c b], a, b \in O, c \in H\}$, as it counts pros and cons criteria.*

- *In π_{SWM}, criteria are not anonymous. Hence weight statements are also needed: $\mathcal{P}(\pi_{SWM}) = \mathcal{P}(\pi_{SM}) \cup \{[c = \alpha], c \in H, \alpha \in \mathcal{S}_H\}$.*

- *In π_M, criteria are anonymous but evaluation statements are needed: $\mathcal{P}(\pi_M) = \mathcal{P}(\pi_{SM}) \cup \{[c : x = \alpha], x \in O, c \in H, \alpha \in \mathcal{S}_O\}$.*

- *In π_{WS}, criteria are not anonymous: $\mathcal{P}(\pi_{WS}) = \mathcal{P}(\pi_M) \cup \{[c = \alpha], c \in H, \alpha \in \mathcal{S}_H\}$.*

A decision model $\pi \in \Pi$ produces a preference relation $\succ_{\pi, P}$ (assumed to be a total order) over the options, given $P \subseteq \mathcal{P}(\pi)$. When P is inconsistent (see Def. 4), $\succ_{\pi, P}$ is empty. Moreover, often, P is *incomplete*, since the DM may not have the ability/time to fully specify the problem. When this is the case, we can use default weights and scores to handle incomplete preference statements (see Ex. 5) , hence the preference order is always complete.

EXAMPLE 5. *(Ex. 4 Cont.) The preference relation derived for the four models π_{SM}, π_{SWM}, π_M and π_{WS} can be put in a unified way. For $\pi \in \Pi$ and $P \in \mathcal{P}(\pi)$ (consistent and possibly incomplete),*

$$a \succ_{\pi, P} b \quad \Leftrightarrow \quad F_{\pi, P}(a, b) > F_{\pi, P}(b, a)$$

where

$$F_{\pi_{SM}, P}(a, b) = |\{c \in H, [a \succ_c b] \in P\}|$$

$$F_{\pi_{SWM}, P}(a, b) = \sum_{c \in H, [a \succ_c b] \in P} \alpha_c^P$$

$$F_{\pi_M, P}(a, b) = \sum_{c \in H} u_c(a)$$

$$F_{\pi_{WS}, P}(a, b) = \sum_{c \in H} \alpha_c^P u_c(a)$$

with

$$\alpha_c^P = \begin{cases} \alpha & \text{if } \exists \alpha \in \mathcal{S}_H \text{ s.t. } [c = \alpha] \in P \\ \text{``average''} & \text{otherwise} \end{cases}$$

(In other words, missing preference information in P regarding weights of criteria is filled by $[c = average]$ (neutral). We assign the numerical weights $\frac{1}{2}$, 1 and 2 to ``weak'', ``average'' and ``strong'' respectively), and

$$u_c(a) = \sum_{d \in O \setminus \{a\}} \Delta_c^P(a, d)$$

$$\Delta_c^P(a, d) = \begin{cases} +3 \text{ if } [a \succ_c d] \in P \text{ and } [c : d = bad] \in P \\ +1 \text{ if } [a \succ_c d] \in P \text{ and } [c : d = bad] \notin P \\ 0 \text{ otherwise} \end{cases}$$

(Missing preference information in P regarding the evaluation of the options on the criteria is filled by the default $+1$ value. We shall not discuss here how figures $+1, +3$ are obtained – see elicitation of intensities of preference, e.g. [6]). Note that utility u_c is computed from differences of intensity of preferences.

The goal of a decision problem, noted G, can be either a *ranking* (from the best option to the worst, as in Ex. 1), or a *selection* of the best option (which is guaranteed to exist since the preference relation is complete). Thus a recommendation is an answer to a given problem G.

DEF. 6. *A comparison statement is of the form $[x \succ y]$ where $x, y \in O$, meaning that x is globally preferred to y.*

DEF. 7. *Two subsets ϕ_1, ϕ_2 of comparison statements are conflicting if there exists $[x \succ y] \in \phi_1$ s.t. $[y \succ x] \in \phi_2$.*

DEF. 8. *If the goal G of the decision problem is a ranking, a recommendation ψ is a subset of comparison statements $[a \succ b]$ which corresponds to a total order over O. If G is the selection of the best option, a recommendation ψ is a subset of comparison statements of the form $\{[a \succ b]$ for all $b \in O \setminus a\}$ for some $a \in O$.*

DEF. 9. *For $P \subseteq \mathcal{P}(\Pi)$ and a subset ψ of preference statements, we define the entailment $\models_\pi$ w.r.t. $\pi \in \Pi$ by*
$$P \models_\pi \{[a_1 \succ b_1], \ldots, [a_q \succ b_q]\} \text{ if } \forall i \in \{1, \ldots, q\} \, [a_i \succ_{\pi, P} b_i].$$

In words, under the decision model π, the consistent preferential information P supports the comparison statements $[a_1 \succ b_1], \ldots, [a_q \succ b_q]$.

EXAMPLE 6 (EX. 1 CONT.). *For $P = \{[d \succ_{c_1} b], [b \succ_{c_2} d], [d \succ_{c_3} b]\}$, we have $P \models_{\pi_{SM}} [d \succ b]$ as $d \succ_{\pi_{SM}, P} b$.*

3.2 Description of the decision models

In order to adapt to different DMs, the DA will use a range of decision models Π, where each model is identified by a set of properties. Such properties correspond to some characteristics of the DM's preferences, corresponding to a set of conditions supporting the use of a given model.

We denote by Q the set of properties that will allow to discriminate among the set of models we consider. For a given model $\pi \in \Pi$, each property can be either satisfied or not. For illustration we will consider the set of properties Q that include: (1) Cardinality of the model (*car*):
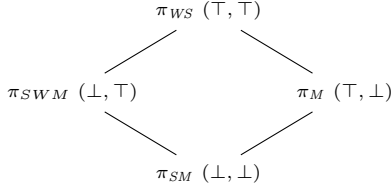
Figure 1: Example of Decision Models

$\pi_{WS}\ (\top, \top)$

$\pi_{SWM}\ (\bot, \top)$ $\pi_M\ (\top, \bot)$

$\pi_{SM}\ (\bot, \bot)$

Figure 2: Structure $\mathcal{Q}$ with three properties

$\pi_{OR}\ (\bot, \top, \top)$ $\pi_{WS}\ (\top, \top, \bot)$

$\pi_{SWM}\ (\bot, \top, \bot)$ $\pi_M\ (\top, \bot, \bot)$

$\pi_{SM}\ (\bot, \bot, \bot)$

it means that the specific difference of performance values makes sense (when this property is not satisfied, only the ordering of options is relevant for comparison). (2) Non-Anonymity of the model (*nan*): it suggests that criteria are not exchangeable (when this property is not satisfied, all criteria are exchangeable). With $Q = \{car, nan\}$, we can describe the four decision models $\pi_{SM}, \pi_{SWM}, \pi_M, \pi_{WS}$.

EXAMPLE 7. *(Ex.2 Cont.) Figure 1 summarizes such models and their description according to the two properties. For instance, π_{SWM} is represented by vector $(\bot, \top)$: the second property (nan) is satisfied (because the weights depend on the criteria), but not the first property (car) as the decision rule does not require cardinality.*

We note that the properties are not supposed to characterize each model (in the sense of axiomatic approaches). For instance, in [13], simple majority is characterized by anonymity, neutrality and monotony. However, in our case, neutrality and monotony are useless to discriminate among the four models [1]. Finally, properties are indeed basically logically independent. However there can be dependencies among them, thereby implying that some combinations of properties is not possible (see Ex. 8).

NOTATION 1. *For $\pi \in \Pi$, let $Q_\pi \subseteq Q$ be the set of properties that decision model π satisfies.*

For instance, $Q_{\pi_{SM}} = \emptyset$ and $Q_{\pi_{WS}} = \{car, nan\}$. Set $\mathcal{Q} = \{Q_\pi , \pi \in \Pi\}$. In our example, $\mathcal{Q} = 2^Q$. But in general, not all subsets of Q correspond to a model. In this case, $\mathcal{Q}$ is assumed to satisfy the following conditions: (i) $\emptyset \in \mathcal{Q}$, there always exists a model fulfilling no property; (ii) if $R \in \mathcal{Q} \setminus \{Q\}$, then $\exists i \in R$ s.t. $R \setminus \{i\} \in \mathcal{Q}$; (iii) If $R, R' \in \mathcal{Q}$, then $R \cap R' \in \mathcal{Q}$. Let us illustrate these properties on a more general situation than Ex. 7.

EXAMPLE 8. *On top of the two properties Cardinality (car) and Non-Anonymity (nan), let us introduce a* veto *property (vet) saying that there is a veto criterion. One can readily see that not all combinations of properties yield to a relevant decision model. Figure 2 shows the set of relevant properties. For instance, the "outranking model" [22] (noted π_{OR}) corresponds to property vector $(\bot, \top, \top)$: it is ordinal but uses criteria weights and veto criteria. On the other hand, property vector $(\bot, \bot, \top)$ has no relevant corresponding model as it satisfies only veto. A similar situation arises for $(\top, \bot, \top)$ and $(\top, \top, \top)$ as a cardinal model (weighted sum) able to represent a veto criterion subsumes to a dictatorial rule (only*

one criterion counts), which is not very interesting and can be represented by π_{OR}. Clearly, the three conditions (i), (ii) and (iii) are satisfied in this example.*

Set $\mathcal{Q}$ is used to guide the navigation among the different models (or associated subsets of properties), depending on the properties that are currently satisfied or contradicted. From (ii), if we take a set R of properties satisfied by a model, then we can remove a property that yields to another set of properties satisfied by a model. By (iii), there exists a model which fulfills only the properties in common of any pair of models. Remark that the second and third property is satisfied by antimatroids and lattices [7], respectively.

3.3 Identifying the decision model of the DA

The DA collects some PI statements P from the DM and then will make inferences. First of all, the DA needs to identify the decision model to use. In fact, given preference statement P, the least specific model (see Def. 10) compatible within P is used by the DA to make assertion, question, challenge, argue in the dialogue (see Axiom 2 in Section 4.3).

Let $\Pi(P) := \{\pi \in \Pi , P \subseteq \mathcal{P}(\pi)\}$ be the set of models compatible with P. In general, several decision models are possible (see example below).

EXAMPLE 9. *(Ex7 Cont.) For our example, if $P = \{[c_2 = very\ strong], [a \succ_{c_2} c], [c \succ_{c_2} b]\}$, then $\Pi(P) = \{\pi_{WSM}, \pi_{WS}\}$ as $P \subseteq \mathcal{P}(\pi_{SWM}), P \subseteq \mathcal{P}(\pi_{WS})$.*

In order to identify the model to use, we introduce the *specificity* of a model. As the elements in Q are basic properties that shall be satisfied by default, the least specific model is the one that satisfies more properties.

DEF. 10. *A model π is* less specific *than π' if $Q_\pi \subseteq Q_{\pi'}$.*

DEF. 11. *Let $\pi[P]$ be the least specific model in $\Pi(P)$. This is the model used by the DA given P.*

In Example 9, $\pi[P]$ is π_{WSM} since it satisfies less properties than π_{WS} as $Q_{\pi_{WSM}} = \{nan\}$ and $Q_{\pi_{WS}} = \{car, nan\}$. More generally, the least specific model is obtained as follows.

EXAMPLE 10. *Given some information P, we can distinguish four cases, summarized in Table 1.*

Intuitively, the notion of specificity also concerns the PI statements that can be used with a model. If decision model π is less specific than π', then π shall use less PI statements, and thus $\mathcal{P}(\pi) \subseteq \mathcal{P}(\pi')$. We strengthen this condition into the following axiom:

AXIOM 1. *Relation Among Models **(RAM)**. Consider three models π_1, π_2, π_{12} such that $R_{12} = R_1 \cap R_2$ where $R_1 = Q_{\pi_1}, R_2 = Q_{\pi_2}, R_{12} = Q_{\pi_{12}}$. Then $\mathcal{P}(\pi_{12}) = \mathcal{P}(\pi_1) \cap \mathcal{P}(\pi_2)$.*

---

[1] Of course, it is always possible to consider more properties in order to describer other types of decision models (interaction among criteria (ruling out additive models), or conditional preferences (leading to CP nets), etc.)
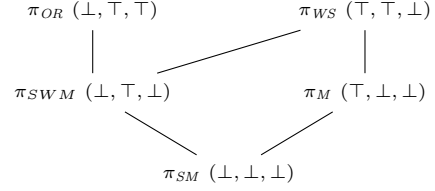
Table 1: Compatible models and least specific model for each type of PI statements.

Form of the statements contained in P	Compatible models	Least specific model
$[a \succ_c b]$	Π	π_{SM}
$[c : x = \alpha]$ and possibly $[a \succ_c b]$	π_M, π_{WS}	π_M
$[c = w_c]$ and possibly $[a \succ_c b]$	π_{SWM}, π_{WS}	π_{SWM}
$[c : x = \alpha]$ and $[c = w_c]$, and possibly $[a \succ_c b]$, $[a \sim_c b]$	π_{WS}	π_{WS}

It is easy to see that **RAM** is satisfied in our running example (Ex. 4).Note that if $R_1, R_2 \in \mathcal{Q}$ then $R_1 \cap R_2 \in \mathcal{Q}$ by condition (iii). This axiom is satisfied in our running example (from Ex. 7 and Figure 1). For instance, with $\pi_1 = \pi_{SWM}, \pi_2 = \pi_M$, we have $\pi_{12} = \pi_{SM}$, $Q_{\pi_{SM}} = Q_{\pi_{SWM}} \cap Q_{\pi_M} = \emptyset$ and $\mathcal{P}(\pi_{SM}) = \mathcal{P}(\pi_{SWM}) \cap \mathcal{P}(\pi_M)$.

Thanks to **RAM**, Definition 11 is well-defined:

LEMMA 1. *Under **RAM**, for any subset P is PI statements, there exists a unique least specific element in $\Pi(P)$.*

LEMMA 2. *For two subsets P, P' of PI statements, if $P \subseteq P'$ then $\pi[P]$ is less specific than $\pi[P']$.*

Proofs are omitted due to space limitations.

4. A FORMAL DIALOGUE MODEL

We have already introduced the two players in the dialogue. The DA has the aim of constructing a solution to a given decision problem. The DM expresses his preferences through feedback and has to be convinced by the solution. Moreover, during the dialogue, the DA constructs a Knowledge Base composed of two parts : $\mathcal{KB}_P \subseteq \mathcal{P}$ containing the Preference Information provided by the DM, and $\mathcal{KB}_\phi$ containing the accepted comparison statements.

EXAMPLE 11. *At the beginning, $\mathcal{KB}_P$ contains all preference statements $[x \succ_{c_i} y]$. In turn 6, $[c_2 = very\ strong]$ is added to $\mathcal{KB}_P$. In turn 8, $[b \succ c]$ is added to $\mathcal{KB}_\phi$*

4.1 Dialogue statements and locutions

We define the *dialogue statements* (Φ) that we need in order to express the different types of information.

DEF. 12. *The* dialogue statements *(Φ) are composed of all comparison statements (see Def. 6) and all preference information (PI) (see Def. 5).*

The different locutions used in our dialogue game are intuitively described below, assuming $\phi \in \Phi$:

- **Assert**(ϕ). It makes possible to put a claim forward.

- **Accept**(ϕ). Used to accept (possibly partially) a claim.

- **Challenge**(ϕ). The challenge requests some statement that can serve as a basis for justifying or explaining ϕ.

- **Question**(ϕ). A question can be used to ask the DM to respond on statement already asserted by the DA. (for instance is it the case that ϕ is true?).

- **Argue**$(\phi,\ p)$ (with $p \subseteq \mathcal{P}$): p is an explanation of ϕ. The link between p and ϕ is set unspecified for the DM, as he does not use in general a model.

- **Contradict**(ϕ) to contradict a previous statement ϕ.

- **Succeed**(ϕ) (such that ϕ is the final recommendation): the DA identifies that it has succeeded in providing a convincing recommendation to the DM.

- **Fail**: the DA acknowledges that it has failed to find a convincing solution to the DM's problem.

4.2 Commitment rules

To capture dialogues between agents, we follow [12, 18] in associating a *commitment store* (CS) with the DM and the DA, which holds the statements and the arguments to which a particular they are *dialectically* committed.

It is however important to stress that the two behave differently: while the DM's one is monotonic, the DA's one can be revised throughout the process. Let $\phi \in \Phi$. In the following table, s stands for the speaker (dm for the DM or da for the DA). The CS is left unchanged with locution **Challenge**.

Assert(ϕ)	$CS(s) = CS(s) \cup \{\phi\}$
Accept(ϕ)	$CS(s) = CS(s) \cup \{\phi\}$
Contradict(ϕ)	$CS(s) = CS(s) \cup \{\neg\phi\}$
Argue(ϕ, p)	$CS(s) = CS(s) \cup \{\phi, p\}$

Note that the locutions **Fail** and **Succeed** mark the end of the dialogue and so will not lead to the updating of the commitment store.

4.3 Dialogue rules

The protocol for our dialogue model is described in Figure 3. Each node in this graph is a locution, except for "Update" (described in detail later), and the outgoing arcs from a node indicate the possible following locutions. A dialogue under this protocol is composed of several iterations. Each iteration starts from node "update", and is organized around an assert(ion) or a question made by the DA, and the feedback of the DM.

In Fig. 3, $\phi_1, \phi_2, \ldots, \phi_8$ are non empty comparison statements, and $p_5, p_7 \subseteq \mathcal{P}$. On top of the previous constraints among locutions, the relevance [17] of the content (dialogue statements) of the moves is constrained (otherwise, the dialogue could easily become meaningless), and the relations among the statements used in successive locutions are specified in the table included in Fig. 3.

For instance, we have $\phi_3 \subseteq \phi_1$ as the DM can challenge only a subpart of what was asserted by the DA.

For the DA, we note that p_5 is formally an explanation of ϕ_5 (i.e. $p_5 \models_{\pi[\mathcal{KB}_P]} \phi_5$). Lastly, we assume that the DM is sure about his preferences and the dialogue will not modify them (they will neither be contradicted nor changed) . This corresponds to the *prescriptive* approach of decision aiding [24]. The aim of the dialogue is to propose a recommendation and a justification to the DM. However, if the DM changes his preferences, the main impact is that statements put in KB_P or KB_ϕ can become wrong later and shall then be revised or removed. Thus when inconsistency arises, the DA may challenge statements in KB_P or KB_ϕ. But, it is outside the scope of this paper to consider this, hence we assume that the dialogue cannot backtrack.

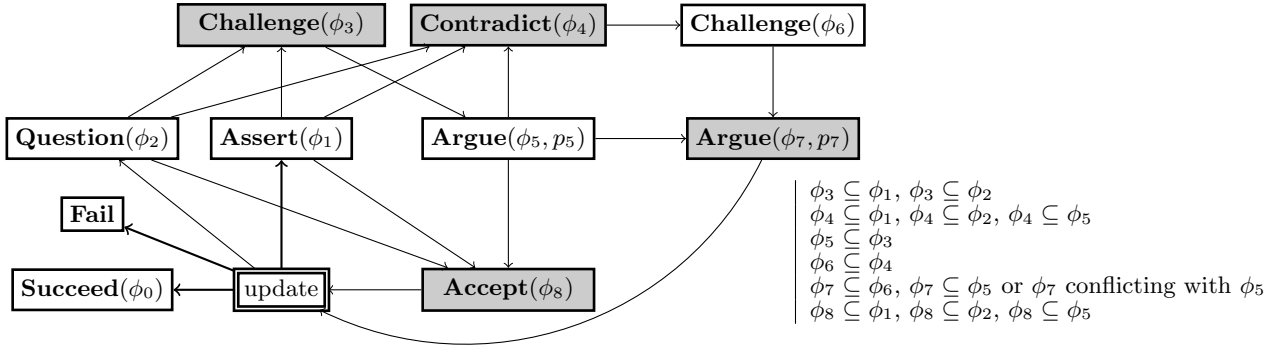Figure 3: Successive speech acts at each iteration (grey nodes are for the DM, white nodes for the DA).

$\phi_3 \subseteq \phi_1$, $\phi_3 \subseteq \phi_2$
$\phi_4 \subseteq \phi_1$, $\phi_4 \subseteq \phi_2$, $\phi_4 \subseteq \phi_5$
$\phi_5 \subseteq \phi_3$
$\phi_6 \subseteq \phi_4$
$\phi_7 \subseteq \phi_6$, $\phi_7 \subseteq \phi_5$ or ϕ_7 conflicting with ϕ_5
$\phi_8 \subseteq \phi_1$, $\phi_8 \subseteq \phi_2$, $\phi_8 \subseteq \phi_5$

The update step.

Node "Update" does not correspond to a speech act. It enables the DA to analyse the exchanges made during last iteration of the dialogue, update the knowledge base and construct the proposal for the next iteration. This is formalized in Axiom **UN**. More precisely, such an axiom presents all cases that can occur in the update node (see Ex 12).

We make several design assumptions. First, we assume that the DA and the DM can use the same statement several times (to allow the DA to update $\mathcal{KB}_P$, $\mathcal{KB}_\phi$ and repropose the same statement. This is for instance the case if the DM agrees with ϕ_1, ϕ_2 or ϕ_5 but not with the argument used). However, the DA is only allowed to propose the same statement more than once if new preference information has been suggested by the DM. Otherwise repetition leads to the protocol ending with a Fail (case (a) below).

AXIOM 2 (UPDATE NODE (**UN**)). *At and after node "Update", the DA behaves as follows:*

(a) *If $CS(dm) \subseteq \mathcal{KB}_P \cup \mathcal{KB}_\phi \cup CS(da)$, then the DA utters* **Fail** *(the DM does not accept new parts of the recommendation, nor does he provide new preferential information. He is not convinced by the arguments of the DA, then the DM and the DA come up with different conclusions with the same preference statements. Hence they cannot agree.);*

(b) *$CS(dm) \cap \mathcal{P}$ is added to $\mathcal{KB}_P$. If $\mathcal{KB}_P$ is inconsistent (Def. 4), the DA makes the speech act* **Fail** *(the information provided by the DM is inconsistent wrt the family of models that the DA can handle.);*

(c) *One identifies the least specific compatible decision model $\pi[\mathcal{KB}_P]$ (see Def. 11). For every $\phi \in CS(dm)$, if $\mathcal{KB}_P \models_{\pi[\mathcal{KB}_P]} \phi$, then ϕ is added to $\mathcal{KB}_\phi$;*

(d) *The recommendation for goal G at current iteration is noted ϕ_c (uniquely defined by Def. 8 and relation $\mathcal{KB}_P \models_{\pi[\mathcal{KB}_P]} \phi_c$). Then the missing commitments for ϕ_c are:*

$$miss(\phi_c) = \phi_c \setminus \mathcal{KB}_\phi \qquad (1)$$

If $miss(\phi_c) = \emptyset$ the DA utters **Succeed**(ϕ_c);

(e) *If $\exists \phi \in CS(dm)$ which contradicts ϕ_c, then the DA makes the speech act* **Assert**$(\neg\phi)$,

(f) *Otherwise: if the current recommendation ϕ_c has not been modified in the update phase, then the DA utters* **Question**(ϕ_1) *with $\phi_1 \subseteq miss(\phi_c)$, or else the DA utters* **Assert**(ϕ_2), *with $\phi_2 \subseteq miss(\phi_c)$.*

Note that this implies that at the first iteration of the protocol, the DA makes the speech act **Assert**(ϕ) with $\phi \subseteq \phi_c$.

From **UN**, the model used by the DM is $\pi[\mathcal{KB}_P]$ and thus the properties that are inferred are $Q_{\pi[\mathcal{KB}_P]}$.

EXAMPLE 12. *(Ex.1 Cont.) In the following we present the different turns of the dialogue. The goal G is a ranking. Superscript "(k)" represents the value at iteration k (for instance, $\mathcal{KB}_P^{(2)}$ is the value of $\mathcal{KB}_P$ at iteration 2). Moreover, when we use the locution statements, we use the labels $\phi_0, \ldots \phi_8$, as in Figure 3, to help the reader to follow the path in the dialogue.*

1st *iteration* – **update**: $\mathcal{KB}_P^{(1)}$ contains all statements $[x \succ_{c_i} y]$, $\mathcal{KB}_\phi^{(1)} = \emptyset$, $\pi[\mathcal{KB}_P^{(1)}] = \pi_{SM}$, $\phi_c^{(1)} = [b \succ a \succ c \succ d]$, $miss(\phi_c^{(1)}) = \phi_c^{(1)}$

(1) DA:**Assert**$(\phi_1^{(1)})$, $\phi_1^{(1)} = \phi_c^{(1)}$
(2) DM:**Challenge**$(\phi_3^{(1)})$, $\phi_3^{(1)} = \{[b \succ a]\}$
(3) DA:**Argue**$(\phi_5^{(1)}, p_5^{(1)})$, $\phi_5^{(1)} = \{[b \succ a]\}$, $p_5^{(1)} = \{[b \succ_{c_3} a], [b \succ_{c_4} a], [b \succ_{c_5} a]\}$
(4) DM:**Contradict**$(\phi_4^{(1)})$, $\phi_4^{(1)} = \{[a \succ b]\}$
(5) DA:**Challenge**$(\phi_6^{(1)})$, $\phi_6^{(1)} = \{[a \succ b]\}$
(6) DM:**Argue**$(\phi_7^{(1)}, p_7^{(1)})$, $\phi_7^{(1)} = \{[a \succ b]\}$, $p_7^{(1)} = \{[a \succ_{c_1} b], [a \succ_{c_2} b], [c_1 = strong], [c_2 = strong]\}$

2nd *iteration* – **update**: $\mathcal{KB}_P^{(2)} = \mathcal{KB}_P^{(1)} \cup \{[c_1 = strong], [c_2 = strong]\}$; $\mathcal{KB}_\phi^{(2)} = \emptyset$, $\pi[\mathcal{KB}_P^{(2)}] = \pi_{WSM}$, $\phi_c^{(2)} = [d \succ a \succ b \succ c]^2$, $miss(\phi_c^{(2)}) = \phi_c^{(2)}$

(7) DA:**Assert**$(\phi_1^{(2)})$, $\phi_1^{(2)} = \{[b \succ c]\}$
(8) DM:**Accept**$(\phi_1^{(2)})$, $\phi_1^{(2)} = \{[b \succ c]\}$: $CS^{(2)}(dm) = \{[b \succ c]\}$

3rd *iteration* – **update**: $\mathcal{KB}_P^{(3)} = \mathcal{KB}_P^{(2)}$; $\mathcal{KB}_\phi^{(3)} = \{[b \succ c]\}$, $\pi[\mathcal{KB}_P^{(3)}] = \pi_{WSM}$, $\phi_c^{(3)} = [d \succ a \succ b \succ c]$, $miss(\phi_c^{(3)}) = \{[d \succ a \succ b]\}$

(9) DM:**Question**$(\phi_2^{(3)})$, $\phi_2^{(3)} = \{[d \succ a]\}$
(10) DA:**Contradict**$(\phi_4^{(3)})$, $\phi_4^{(3)} = \{[a \succ d]\}$

---

[2]In particular $d \succ_{\pi_{WSM}, \mathcal{KB}_P^{(2)}} a$ as $\alpha_{c_1} = \alpha_{c_2} = 2$ and $\alpha_{c_3} = \alpha_{c_4} = \alpha_{c_5} = 1$.

(11) DM:**Challenge**$(\phi_6^{(3)})$, $\phi_6^{(3)} = \{[a \succ d]\}$

(12) DA:**Argue**$(\phi_7^{(3)}, p_7^{(3)})$, $\phi_7^{(3)} = \{[a \succ d]\}$, $p_7^{(3)} = \{[a \succ_{c_3} d], [a \succ_{c_4} d], [a \succ_{c_5} d], [c_3 : d = bad], [c_4 : d = bad], [c_5 : d = bad]\}$

$\underline{4^{rd}\ \textit{iteration}}$ – **update**: $\mathcal{KB}_P^{(4)} = \mathcal{KB}_P^{(2)} \cup \{[c_4 : d = bad]\}$; $\mathcal{KB}_\phi^{(4)} = \{[b \succ c]\}$, $\pi[\mathcal{KB}_P^{(4)}] = \pi_{WS}$, $\phi_c^{(4)} = [a \succ b \succ c \succ d]$, $miss(\phi_c^{(4)}) = \{[a \succ b], [c \succ d]\}$

Let us explain why $c \succ d$. For the computation of Δ_c, we have for instance $\Delta_2^P(d,c) = 1$ and $\Delta_3^P(c,d) = 3$. Hence $u_1(c) = \Delta_1(c,a) + \Delta_1(c,b) + \Delta_1(c,d) = 1$, $u_2(c) = 0$, $u_3(c) = 4$, $u_4(c) = 5$, $u_5(c) = 3$, and $u_1(d) = 3$, $u_2(d) = 3$, $u_3(d) = 0$, $u_4(d) = 0$, $u_5(d) = 0$. Moreover, $F_{\pi_{WS},P}(c,d) = \alpha_1^P u_1(c) + \alpha_2^P u_2(c) + \alpha_3^P u_3(c) + \alpha_4^P u_4(c) + \alpha_5^P u_5(c) = 14$, $F_{\pi_{WS},P}(d,c) = 12$ so that $c \succ_{\pi_{WS},P} d$.

(13) DA:**Assert**$(\phi_1^{(4)})$, $\phi_1^{(4)} = \{[c \succ d]\}$

(14) DM:**Accept**$(\phi_8^{(4)})$, $\phi_8^{(4)} = \{[c \succ d]\}$): $CS^{(4)}(dm) = CS^{(2)}(dm) \cup \{[c \succ d]\}$

$\underline{5^{rd}\ \textit{iteration}}$ – **update**: $\mathcal{KB}_P^{(5)} = \mathcal{KB}_P^{(4)}$; $\mathcal{KB}_\phi^{(5)} = \{[b \succ c], [c \succ d]\}$, $\pi[\mathcal{KB}_P^{(5)}] = \pi_{WS}$, $\phi_c^{(5)} = [a \succ b \succ c \succ d]$, $miss(\phi_c^{(5)}) = \{[a \succ b]\}$

(15) DA:**Question**$(\phi_2^{(5)})$, $\phi_2^{(5)} = \{[a \succ b]\}$

(16) DM:**Accept**$(\phi_8^{(5)})$, $\phi_8^{(5)} = \{[a \succ b]\}$, $CS^{(5)}(dm) = CS^{(4)}(dm) \cup \{[a \succ b]\}$

$\underline{6^{rd}\ \textit{iteration}}$ – **update**: $\mathcal{KB}_P^{(6)} = \mathcal{KB}_P^{(4)}$; $\mathcal{KB}_\phi^{(6)} = \{[a \succ b \succ c \succ d]\}$, $\pi[\mathcal{KB}_P^{(6)}] = \pi_{WS}$, $\phi_c^{(6)} = \{[a \succ b \succ c \succ d]\}$, $miss(\phi_c^{(6)}) = \emptyset$

(17) DA:***Success***$(\phi_0^{(6)})$, $\phi_0^{(6)} = \{[a \succ b \succ c \succ d]\}$

In this example, we start with model π_{SM} at the first iteration. Then model π_{SWM} is used at the second iteration due to statements $[c_1 = strong]$, $[c_2 = strong]$. Lastly at iteration 4, π_{WS} is used due to statements $[c_3 : d = bad], [c_4 : d = bad], [c_5 : d = bad]$. The inference of the comparison among options is consistently constructed even though the model is changing, thanks to the relation between the models and the related properties.

5. TERMINATION OF THE DIALOGUE

At each new iteration of the dialogue, there are two possible end states: success (acceptance by the DM of a recommendation), or a failure (the DA is not able to find a proposal with an explanation that convinces the DM).

PROPOSITION 1. *Under **UN**, the length of the dialogue resulting from the protocol is at most:*

$$7\,|\overline{P}| + 2\,|O|\,(|O| - 1) + 1$$

where $\overline{P}$ is the knowledge base of the DM.

The size of $\overline{P}$ depends on the number of criteria. One can easily derive bounds of $|\overline{P}|$ from the type of models that the DM is expected to follow.

COROLLARY 1. *Under **UN**, the protocol terminates.*

Termination requires very few assumptions. However, as we shall see now, obtaining guarantees on the quality of the outcome is much more demanding.

6. OUTCOMES OF THE DIALOGUE

The DA is deemed to be an automatic agent following some rationality postulates (e.g. axiom **UN**). On the other hand, the DM is an individual and has more freedom of action in the dialogue. However, we show in this section that if the DM is representable by a model contained in the set Π of models, then the dialogue necessarily terminates with a **Succeed**, the option that results from the dialogue is among the best options for the DM, and the properties that the DA guesses are correct (but the DA may not have guessed *all* properties – this depends on the length of the dialogue). In particular, if the dialogue ends with a failure, this means that the DM is not representable by a model in Π. In order account for this, we should make some assumptions of the consistency of both the DA and DM: in particular, the DM must accept a statement if he agrees with the explanation provided by the DA.

We first strengthen the constraint of the explanation given by the DA, following a *data-based explanation approach* [9].

AXIOM 3 (EXPLANATION IN ARGUE (**EA**)). *Consider an agent (DA or DM) having preferences P and using model π.*

*For the agent to utter **Argue**$([x \succ y], p)$, p is the set of all statements of the form $[x \succ_c y]$, $[y \succ_c x]$, $[c = w_c]$, $[c : x = \alpha]$ and $[c : y = \alpha]$ belonging to P.*

*For the agent to utter **Argue**(ϕ, p), p is the union of all p statements appearing in **Argue**$([x \succ y], p)$, for all elements $[x \succ y]$ of ϕ. In particular, $p \models_\pi \phi$.*

We consider the case where DM is represented by preference information $\overline{P}$ and user model $\overline{\pi} := \pi[\overline{P}]$ (Def. 11). In our running example, we have $\overline{P}$ contains all statements of the form $[x \succ_{c_i} y]$, plus $[c_1 = strong]$, $[c_2 = strong]$ and $[c_4 : d = bad]$. Moreover, $\overline{\pi} = \pi_{WS}$.

We can illustrate axiom **EA** from Ex. 12. At turn (3), the DA argues $\{[b \succ a]\}$, by the explanation $\{[b \succ_{c_3} a], [b \succ_{c_4} a], [b \succ_{c_5} a]\}$. The explanation indeed contains all statements in $\mathcal{KB}_P^{(1)}$ that are related to the comparison $[b \succ a]$. The same holds for the other speech acts **Argue** used throughout the dialogue (see turns (6), (12)).

AXIOM 4 (CONSISTENCY FOR THE DM (**C**)). *We assume that $\overline{P}$ is consistent. If the DA utters **Argue**(ϕ_5, p_5) in the protocol, then the next speech act is:*

(α) *The DM utters **Contradict**(ϕ_4) iff there exists ϕ_4' s.t. $\phi_4 \subseteq \phi_5$, $\overline{P} \models_{\overline{\pi}} \phi_4'$ and ϕ_4 is conflicting with ϕ_4';*

(β) *The DM utters **Accept**(ϕ_8) iff $\phi_8 \subseteq \phi_5$, $\overline{P} \models_{\overline{\pi}} \phi_5$ and $p_5 \models_{\overline{\pi}} \phi_5$ (the DM would obtain the same conclusion with his preferences and also the same explanation).*

(γ) *Otherwise, the next move of the DM is **Argue**(ϕ_7, p_7), where $\phi_7 \subseteq \phi_5$, $p_7 \subseteq \overline{P}$, $p_7 \models_{\overline{\pi}} \phi_7$ and $p_7 \not\subseteq p_5$ (the DM agrees on ϕ_7 but provides a more specific explanation).*

EXAMPLE 13. *In Turns (4), the DM asserts a statement that is exactly the opposite to the statement argued just before by the DA, which fulfills axiom **C**.*

LEMMA 3. *Let $\pi \in \Pi$ and $P \in \mathcal{P}(\pi)$. If **Argue**(ϕ, p) is used (with $p \subseteq P$), then for every $p' \supseteq p$ with $p' \subseteq P$, p' consistent and $p' \in \mathcal{P}(\pi)$, then $p' \models_\pi \phi$.*

LEMMA 4. *Let $P \subseteq \mathcal{P}$. $\mathcal{Q}(P) = \{R \supseteq Q_{\pi[P]} , R \in \mathcal{Q}\}$.*

PROPOSITION 2. *Assume that **RAM, EA, UN** and **C** are satisfied. Let $\overline{R} = Q_{\overline{\pi}}$. Assume that the knowledge base of the DA at the start of the dialogue is included in $\overline{P}$. Then:*

- *The dialogue terminates with **Success***;

- *The dialogue stops with properties $R \in \mathcal{Q}$, and $R \subseteq \overline{R}$ (the properties guessed by the DA are correct);*

- *at the end, the recommendation provided by the DA is $\succsim_{\overline{\pi},\overline{P}}$, and the DM agrees with it.*

PROOF. In an iteration of the protocol, the knowledge bases are $\mathcal{KB}_P$ and $\mathcal{KB}_\phi$. Axiom **UN** determines the model and thus the properties R corresponding to the preference information $\mathcal{KB}_P$ collected so far by the DA: R is the smallest element of $\mathcal{Q}(\mathcal{KB}_P)$ (w.r.t. $\subseteq$). Let π be the model associated to R (i.e. with $Q_\pi = R$). Hence $\pi = \pi[\mathcal{KB}_P]$ and $R = Q_{\pi[\mathcal{KB}_P]}$. By the statement of the proposition, $\overline{R}$ is the smallest element of $\mathcal{Q}(\overline{P})$ w.r.t. $\subseteq$. By definition of $\mathcal{KB}_P$, we have $\mathcal{KB}_P \subseteq \overline{P}$. Clearly, by **RAM**, we have $\mathcal{Q}(\mathcal{KB}_P) \supseteq \mathcal{Q}(\overline{P})$ and thus $\overline{R} \in \mathcal{Q}(\mathcal{KB}_P)$ (as $\overline{R} \in \mathcal{Q}(\overline{P})$).

By Lemma 4, we have $\mathcal{Q}(\mathcal{KB}_P) = \{R' \supseteq R , R' \in \mathcal{Q}\}$. Hence property $\overline{R} \in \mathcal{Q}(\mathcal{KB}_P)$ implies that $R \subseteq \overline{R}$.

Assume by contradiction that the dialogue ends by **Fail**. By **UN**, a Fail is obtained only when the last move of the DM is a **Argue**(ϕ_7, p_7). This speech act was a respond to statement ϕ (in **Argue**(ϕ_1), **Question**(ϕ_2) or **Argue**(ϕ_5, p_5)), by the DA. We assume that ϕ is supported by p, with $p \subseteq \mathcal{KB}_P$, i.e. $p \models_\pi \phi$ by Axiom **EA**. There are two cases:

Case 1: $p_7 \subseteq \mathcal{KB}_P$ – case **UN-(a)**: the DM did not provide any new preference information. As the DM argued, he did not agree with **Argue**(ϕ, p) made by the DA.

In the case **UN-(a)**, we have $CS(dm) \subseteq \mathcal{KB}_P \cup \mathcal{KB}_\phi \cup CS(da)$. This implies that the DM arrives at the same conclusions as the DA.

We conclude that ϕ and ϕ_7 cannot be conflicting (see Def. 7).

The DM could not have used speech **Contradict**(ϕ_4) since then ϕ_4 (which contradicts a statement commited by the DA) would belong to $CS(dm)$, and thus $CS(dm) \not\subseteq \mathcal{KB}_P \cup \mathcal{KB}_\phi \cup CS(da)$, which contradicts **UN-(a)**.

Hence in the last iteration of the dialogue, there is necessarily the speech act **Argue**(ϕ_5, p_5) by the DA, and then later the speech act **Argue**(ϕ_7, p_7) by the DM, with $\phi_7 \subseteq \phi_5$.

As the DM didn't contradict **Argue**(ϕ_5, p_5) (by the DA), the DM agrees with ϕ_5 (see **C-**α). Hence $\overline{P} \models_{\overline{\pi}} \phi_5$. Now, as the DM didn't accept **Argue**(ϕ_5, p_5) (by the DA), we have $p_5 \not\models_{\overline{\pi}} \phi_5$.

Furthermore, as the DM made speech act **Argue**(ϕ_7, p_7), we have $\phi_7 \subseteq \phi_5$, $p_7 \subseteq \overline{P}$, $p_7 \models_{\overline{\pi}} \phi_7$ and $p_7 \not\subseteq p_5$. Then, $p_7 \subseteq \mathcal{KB}_P$ as $p_7 \subseteq CS(dm)$ and $CS(dm) \subseteq \mathcal{KB}_P \cup \mathcal{KB}_\phi \cup CS(da)$.

To sum-up, we have

$$p_5 \models_\pi \phi_5 , \ p_5 \not\models_{\overline{\pi}} \phi_5 , \ \overline{P} \models_{\overline{\pi}} \phi_5,$$
$$p_7 \models_{\overline{\pi}} \phi_7 , \ \phi_7 \subseteq \phi_5,$$
$$p_7 \not\subseteq p_5 , \ p_7 \subseteq \mathcal{KB}_P , \ p_5, p_7 \in \mathcal{P}(\pi)$$

From **EA**, p_5 (resp. p_7) contains all statements in $\mathcal{KB}_P$ related to ϕ_5 (resp. ϕ_7). As $\phi_7 \subseteq \phi_5$, it is not possible $p_7 \not\subseteq p_5$. Hence a contradiction is raised.

Case 2: $\mathcal{KB}_P$ is inconsistent (after p_7 has been added to $\mathcal{KB}_P$) – case **UN-(b)**. This is not possible as $\mathcal{KB}_P$ contains

only the preference information provided by the DM (i.e. $\mathcal{KB}_P \subseteq \overline{P}$), and the preference information $\overline{P}$ is consistent and thus any subset is also consistent. Hence the dialogue cannot end by **Fail**.

As the dialogue terminates (see Proposition 1), it necessarily terminated by a Success. By **UN**, a Success occurs when the DM has accepted (in one or several times) the recommendation of the DA for goal G that the DA can derive from $\mathcal{KB}_P$ and π. By **C-**(β), the DM accepts a statement only if it is entailed by his preferences. Hence the DM agrees with the recommendation of the DA for goal G and it is necessarily the final recommendation. $\square$

This is our main result: it shows for instance that if the protocol returns a single recommended option, then this option is indeed amongst the DM's most preferred options.

7. RELATED WORK AND CONCLUSION

Recommender systems have developed very sophisticated techniques and algorithms, with the DM feedback being as a vital component allowing to produce better recommendations. However, the case of multi-criteria recommendation remains challenging: it was identified as an emerging topic in the survey of [1] and is still recognized as such in the Recommender Systems Handbook [21]. A problem arising in this context is that it opens a wide range of possible models to account for the DM's preferences [6]. And in that case, the feedback of the DM may reveal preferential information that require more than a simple adjustment of a parameter in a predefined model. For instance in [20], a weighted sum model is used. For a given criterion, its weight is initialized by a default value, and is then multiplied by a factor if the user critiques this criterion (the critique proves the user put more importance on this criterion). While our approach is close in spirit, we instead show in this paper how the feedback of the DM can be exploited so as to perform adaptive selection of preference models.

Dialectical models of interaction have gained tremendous popularity in recent years in the multiagent community. Many protocols have been put forward, to tackle different types of interaction [26]. It is clear that these protocols offer a greater expressivity than simple feedback (since recommendations can be challenged and justified, as illustrated here). Recently, an emphasis has been put on proving properties of such dialectical models, see *e.g.* [3, 10]. Our paper follows this trend of research and studies a type of interaction whose specificities have seldom been studied. Indeed, while the link between decision-making and argumentation has been investigated over a number of years [2, 8, 11, 15, 23], the decision-aiding setting itself has been little studied, and the little reported work [16] does not go as far as we do in capturing the process of exploring possible decision models.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state of the art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, 2005.

[2] L. Amgoud and H. Prade. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3-4):413–436, 2009.

[3] E. Black and A. Hunter. Executable logic for dialogical argumentation. In *Proc. of ECAI*, pages 15–20, 2012.

[4] C. Boutilier, R. Patrascu, P. Poupart, and D. Schuurmans. Constraint-based optimization and utility elicitation using the minimax decision criterion. *Artif. Intel.*, (170):686–713, 2006.

[5] C. Boutilier, R. Zemel, and B. Marlin. Active collaborative filtering. In *19th Conf. on Uncertainty in AI (UAI-07)*, 2003.

[6] D. Bouyssou, T. Marchant, M. Pirlot, P. Perny, A. Tsoukiàs, and P. Vincke. *Evaluation and decision models: a critical perspective*. Kluwer Academic, 2000.

[7] B. Davey and H. Priestley. *Introduction to Lattices and Orders*. Cambridge University Press, 1990.

[8] J. Fox and S. Parsons. Arguing about beliefs and actions. In *Applications of uncertainty formalisms*. Springer-Verlag, 1998.

[9] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *CSCW*, pages 241–250, 2000.

[10] A. Hunter. Analysis of dialogical argumentation via finite state machines. In *Proc. SUM*, 2013.

[11] A. Kakas and P. Moraitis. Argumentation based decision making for autonomous agents. In *Proc. AAMAS*, 2003.

[12] J. D. MacKenzie. Four dialogue systems. *Studia Logica*, pages 567–583, 1990.

[13] K. O. May. A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision. *Econometrica*, 20(4):680–684, 1952.

[14] L. McGinty and B. Smyth. Adaptive selection: An analysis of critiquing and preference-based feedback in conversational recommender systems. *Int. J. Elec. Comm.*, 11(2):35–57, 2007.

[15] J. Müller and A. Hunter. An argumentation-based approach for decision making. In *Proc. ICTAI*, 2012.

[16] W. Ouerdane, N. Maudet, and A. Tsoukiàs. Dealing with the dynamics of proof-standard in argumentation-based decision aiding. In *Proc. ECAI*, pages 999–1000, 2010.

[17] S. Parsons, P. McBurney, E. Sklar, and M. Wooldridge. On the relevance of utterances in formal inter-agent dialogues. In *Proc. AAMAS*, 2007.

[18] H. Prakken. Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, (21):163–188, 2006.

[19] R. Price and P. Messinger. Optimal recommendation sets: Covering uncertainty over user preferences. In *20th National Conf. on AI (AAAIÕ05)*, 2005.

[20] J. Reilly, J. Zhang, L. McGinty, P. Pu, and B. Smyth. Evaluating compound critiquing recommenders: a real-user study. In *ACM Conf. on Electronic Commerce*, 2007.

[21] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[22] B. Roy. How outranking relations helps multiple criteria decision making. In J.-L. Cochrane and M. Zeleny, editors, *Multiple Criteria Decision Making*, pages 179–201. University of South California Press, Columbia, 1973.

[23] P. Tolchinsky, S. Modgil, U. Cortes, and M.Sanchez-Marre. Cbr and argument schemes for collaborative decision making. In *Proc. COMMA*, 2006.

[24] A. Tsoukiàs. On the concept of decision aiding process. *Annals of Operations Research*, pages 3 – 27, 2007.

[25] P. Viappiani, B. Faltings, and P. Pu. Preference-based search using example-critiquing with suggestions. *J. Artif. Intell. Res.*, (27):465–503, 2006.

[26] D. Walton and E. Krabbe. *Commitment in Dilaogue : Basic conceptions of Interpersonal Reasoning*. State University of New York Press, 1995.

Towards Automating Decision Aiding Through Argumentation

WASSILA OUERDANE[a], YANNIS DIMOPOULOS[b*], KONSTANTINOS LIAPIS[c] and PAVLOS MORAITIS[d]

[a] *LGI, Ecole Centrale Paris, Paris, France*
[b] *Department of Computer Science, University of Cyprus, Nicosia, Cyprus*
[c] *Department of Economic and Regional Development, Panteion University, Athens, Greece*
[d] *LIPADE, Paris Descartes University, Paris, France*

ABSTRACT

Decision aiding can be abstractly described as the process of assisting a user/client/decision maker by recommending possible courses of his action. This process has to be able to cope with incomplete and/or inconsistent information and must adapt to the dynamics of the environment in which it is carried out. Indeed, on the one hand, complete information about the environment is almost impossible, and on the other hand, the information provided by the user is often affected by uncertainty; it may contains inconsistencies and may dynamically be revised because of various reasons. The aim of this paper is to present a model of the decision aiding process that is amenable to automation. The main features of the approach is that it models decision aiding as an iterative defeasible reasoning process, and it uses argumentation for capturing important aspects of the process. More specifically, argumentation is used for representing the relations between the cognitive artefacts that are involved in decision aiding, as well as for modelling the artefacts themselves. In modelling the cognitive artefacts, we make use of the notion of argument schemes and specify the related critical questions. More specifically, the work reported here aims at initiating a systematic study of the use of argumentation in future decision aiding tools. Our ambition is twofold: (i) enhance decision support capabilities of an analyst representing explicitly and accountably the reasons for which he recommend a solution for a decision maker and (ii) enhance decision support capabilities of an (semi) automatic device to handle (at least partially) the dialogue with the user. Copyright © 2011 John Wiley & Sons, Ltd.

KEY WORDS: argumentation; decision aiding; artificial intelligence

1. INTRODUCTION

Decision analysis (Belton and Stewart, 2002; Bouyssou *et al.*, 2000; French, 1988) is concerned with the process of providing decision support to 'clients' who feel unable to handle alone a problem situation. We call such an activity 'decision aiding'. Decision aiding is a process characterized by the emergence of cognitive artefacts, resulting from the interaction between the 'client' and the 'analyst'. The decision analyst and the client are engaged in an iterative process, where the analyst attempts, through successive steps of interaction with the client, to obtain a better understanding of the problem the client is facing. To be able to cope with the complexity of both the real world and the needs of the client, the analyst needs to make assumptions and reason as if these assumptions were true in the world. The recommendations, which are the outcomes of the decision aiding process (DAP), are subject to the client validation. Rejection of the recommendations means that some of the assumptions made by the analyst are false and must be retracted.

On the contrary, systems that aim at assisting people in decision making help the user to shape a problem situation, formulate a problem and possibly try to establish a viable solution to it. Decision theory and Multiple-Criteria Decision Analysis have established the theoretical foundation upon which many decision support systems have blossomed. These approaches (and the formal tools coming along with them) have focused for a long time on how a 'solution' should be established. But it is clear that the process involves many other aspects that are handled more or less formally by the analyst. For instance,

- The problem of *accountability* of decisions is almost as important as the decision itself. The decision maker should then be convinced by a proper

*Correspondence to: Department of Computer Science, University of Cyprus, Nicosia, Cyprus.
E-mail: yannis@cs.ucy.ac.cy

explanation that the proposed solution is indeed the best (see Belton and Stewart, 2002; Bouyssou *et al.*, 2000).

- It should be possible, for the client, to *refine*, or even *contradict*, a given recommendation. Decision–support processes often need to be constructive, in the sense that a user may revise the assumptions or other aspect of the problem description when the potential solutions and their implications become explicit.

Nowadays, decision-aiding situations are pervasive: they can occur in situations where the role of the analyst is taken by a non-expert, in some extreme cases even by an automatic tool. For instance, consider the following scenarios:

- Ann is not an experienced analyst, but she has good knowledge of some decision–support tools that she herself used quite often. She would like to help Bob to make a decision regarding some public policy investment. In this situation, Ann may find useful to have the support of a tool that would provide her with explicit explanations, justifications and possible replies that could occur in the course of an interaction with her 'client'. Similarly, such a system could be used for the non-expert analyst to practice and simulate some virtual interactions with a client.
- Bob is purchasing items on the Internet. He has to choose among a selection of 150 digital cameras on a commercial website (too many to be examined exhaustively). Bob first provides some preferential information to the system. On the basis of the responses of Bob to these questions, the *recommender system* selects a specific model. Bob, not fully satisfied or convinced by the recommendation, would like to interact with the system, at least to gain a better understanding of the reasons underlying this. Such needs has been identified by mainstream recommender systems (Chen and Pu, 2007) but is only very simply addressed. For instance, it is now possible to check *why* a given item has been recommended by Amazon and to contradict the relevance of a certain purchase act for forthcoming recommendations.

The aforementioned scenarios mean that several aspects usually delegated to the human analyst should (in these situations) ideally be handled by the decision–support system. The task is ambitious: in a 'human-to-human' interaction—even though the

dialogue is possibly supported by standard protocols (as in the case of constructing a value or an utility function or assessing importance parameters) that fix some explicit formal rules on how such a process can be conducted—the dialogue is handled through typical human interaction. A tool should be able to *structure* the dialogue on a formal basis in order to be able to control and assess what the artefact concludes as far as the user preference models are concerned and what type of recommendations (if any) is going to reach. In short, we need on the one hand some formal theory about preferences (and this is basically provided by decision analysis), and on the other hand some formal language enabling to represent the dialogue, to explain it, to communicate its results, to convince the user/decision maker that what is happening is both theoretically sound and operationally reasonable.

Although, in the decision analysis literature, there was until recently very little attention to the use of decision theories and decision aiding methodology when the interaction occurs between a human (a user) and an automatic device (see Klein, 1994 for a noticeable exception), the recent surge of automatic decision aiding tools on the Internet (recommender systems) have motivated a great deal of research. For instance, there are studies on the impact of higher levels of interaction with the user or explanation capabilities on the efficiency of the recommendations (Pu and Chen, 2007). Because of the context however, only very simple interactions and models of preferences are envisaged (a typical consumer is not prepared to enter in a long preference elicitation process or to discuss endlessly the benefits of a given option as opposed to another one).

At the same time, there is a long tradition in artificial intelligence (AI), going back to the early work of Simon, to challenge some assumptions of decision theory models or to emphasize their limits in certain circumstances. Stimulated by the objective to design agents capable of autonomous decision-making abilities (think of a robot exploring planet Mars), AI researchers pointed out the need to deal with missing or incomplete information, to revise some objectives to adapt to the new contexts and so on. In particular, the *knowledge representation* trend of AI has greatly contributed to challenge and question the rather crude 'utility' models used in decision theory. Indeed, one of the key distinctive ingredient of many AI-based approaches is to represent decision making in terms of 'cognitive attitudes' (as exemplified in the famous Belief-Desire-Intention paradigm) (Dastani *et al.*, 2005; Doyle and Thomason, 1999), instead of mere

utilities (as already elicited by the analyst). This change of perspective paved the way for more *flexible* decision-making models: goals may change with circumstances, and understanding these underlying goals offers, for instance, the opportunity to propose alternative actions. The approach is attractive as it offers a natural and powerful way to specify agents' preferences. This is because it naturally caters for partial specification of preferences and makes explicit many aspects that are usually somewhat hidden in decision models.

A second, very influential contribution of AI, related to the previous point, has been the development of techniques for reasoning in the presence of conflicting (possibly heterogeneous) information. As the DAP is based on retractable assumptions, the formal modelling language to be used must be a non-monotonic one. Among the several possibilities, it seems that argumentation is particularly well suited for this task. Indeed, recently, following some early works (Bonet and Geffner, 1996; Fox and Parsons, 1997), several models have been put forward in the AI community that makes use of argumentation techniques (Amgoud, 2009; Amgoud *et al.*, 2005; Atkinson *et al.*, 2006; Labreuche, 2006; Ouerdane *et al.*, 2007, 2009) to tackle decision problems. Such approaches have identified a variety of argument structures allowing to highlight the benefits of argumentation for decision: expressiveness and explicit representation of reasoning steps. Thus, they have greatly extended our understanding of the construction of argument for action or decision.

Moreover, the use of argumentation in decision support systems has been ever increasing. Such systems aim at assisting people in decision making. The need to introduce arguments in such systems has emerged from the demand to justify and to explain the choices and the recommendations provided by them. Together with this, other needs have motivated the use of argumentation, such as dealing with incomplete information, qualitative information and uncertainty (Amgoud and Prade, 2006; Fox *et al.*, 1993; Parsons and Greene, 1999). Such systems deal with different contexts and applications, which may involve very different types of decision makers, from experts (medical domains) (Atkinson *et al.*, 2006; Shankar *et al.*, 2006) to potential buyers (recommender systems) (Chesñevar and Maguitman, 2004; Chesñevar *et al.*, 2004, 2006) or simple citizens (public debate) (Atkinson, 2006; Morge, 2004), and even largely autonomous pieces of software that should act on behalf of a user (multi-agent systems) (Kakas

and Moraitis, 2003; Parsons and Jennings, 1998; Sillince, 1994).

It is important to note that many of the aspects of decision-support systems discussed previously touch upon issues that have been long identified in system design. Indeed, it was about four decades ago when researchers such as Rittel (Rittel and Webber, 1973) came to realize that in order to tackle ill-defined problems (as opposed to the well-defined problems of science), an 'argumentative approach' was needed. This initiated the *design rationale* movement that advocates the thesis that 'to understand why a system design is the way it is, we also need to understand how it could be different, and why the choices which where made are appropriate' (MacLean *et al.*, 1989). The rationale for a system describes the decision that have taken the possible alternatives that were considered and the pros and cons of these alternatives. Not surprisingly, many design rationale systems, starting with the early IBIS (Issue-Based Information System) (Kunz and Rittel, 1970) system to the more recent SEURAT (Software Using RATionale system) (Burge and Brown, 2008), are based on some form of argumentation.

The work presented here is in accordance with previous studies that employ argumentation in decision making but from a different perspective. First, the approach described here is not based on cognitive attitudes and the underlying motivation of the decision maker, but it relies on information provided by the decision maker during a Decision Aiding Process (DAP). Second, decision aiding is understood as the process of constructing and revising cognitive artefacts, which gradually transform an abstract problem description to an invocation of a concrete decision support tool. Therefore, automating the DAP amounts to automating the construction of these cognitive artefacts taking also into account the defeasible character of the process. Argumentation is the language that captures the interdependencies between these artefacts and controls the overall process.

The specific approach to artefact construction that is taken in this work is not intended to provide a fully automated and general approach to decision aiding, as it is limited in several ways. Initially, it is restricted to the modelling of specific cognitive artefacts involved in the Decision Aiding Process, leaving out highly abstract cognitive tasks such as the representation of a problem situation. Furthermore, the construction of the artefacts is a process where *predefined* 'components' of a decision–support system are put together to make a meaningful whole. Argumentation maintains a *high-level control* of this synthetic process and enables, instead of a static 'composition' of the

elements of the system at design time, a more dynamic one that can be decided at run-time, through the interaction with the environment or the user. Therefore, instead of building a rigid decision aiding tool, argumentation offers the possibility of delivering different, context-dependent versions of such a tool.

Furthermore, we note that even within the limited scope of the method that is described in this work, it is not always easy or even possible to completely automate this process because on the one hand, it is not always easy to model the decision aiding methodology, and the expertise of the analyst in the decision problem is considered. On the other hand, it is also not always easy or obvious to identify all needs and information necessary to fully meet the expectations of the user. In such cases, the approach can be seen as a method of building tools that support the DAP, which may help both the analyst and the user to share a common representation of the problem and the proposed solutions.

The proposed approach aims at facilitating the development of automatic decision devices and the improvement of decision support and recommender systems. In fact, the diffusion of Web-based services pushed the development of online decision support and decision support and recommender systems for a large variety of applications (e-commerce, e-voting, e-services, semantic Web, etc.). Such tools have to combine traditional decision making methods with flexible reasoning procedures allowing to handle the large variety of tasks required, to be adaptable to the changing environment where they operate and to perform self-improvement. Therefore, automating the DAP, whenever this is possible, by using argumentation is a first step towards meeting these needs. Additionally, argumentation can be used to provide design rationale information to future users and developers.

In summary, this paper studies two different ways of employing argumentation in decision aiding. First, we show how the relation between different artefacts of the DAP can be modelled using the framework of Kakas and Moraitis (2003), which is dynamic in the sense that the arguments, and their strength depend on the particular context that the decision maker (or agent) finds himself, thus allowing the agent to adapt his decisions in a changing environment. The decision aiding theories can be easily implemented directly from their declarative specification in the Gorgias system (Gorgias, 2002) for this framework. We focus on the inferences that can be drawn by argumentation and the way these inferences can be retracted in the light of the new information, capturing in this manner the dynamics of the DAP.

Second, we investigate how argumentation can model crucial aspects of each artefact of the DAP. More specifically, we study how *argument schemes* can be developed and used in the DAP. Argument schemes are presented as general inference rules whereby given a set of premises, a conclusion can be drawn. However, such schemes are not deductively strict because of the defeasible nature of arguments (Norman *et al.*, 2003; Walton, 1996, 2005). The schemes allow for arguments to be represented within a particular context and take into account the fact that the underlying reasoning may be altered in the light of new evidence or exceptions to rules.

The paper is structured as follows. In Section 2, first we introduce the concept of the DAP and the cognitive artefacts it produces. Then, we explain, by means of an example, what is missing in such a process. In Section 3, we briefly introduce the argumentation framework we use and show how it can capture the relations between the cognitive artefacts. Section 4 presents how the notion of arguments schemes, and their related critical questions, can be used to represent the steps of a multicriteria evaluation process. Finally, Section 5 provides the conclusion.

2. THE DECISION AIDING PROCESS

Decision aiding is an activity occurring in the everyday life of almost everybody. In this paper, we are interested in that particular type of decision aiding where formal and abstract languages are used (different decision theories and approaches). A DAP is a particular type of decision process involving at least two actors: a client, who himself is involved in at least one decision process (the one generating the concern for which the aid is requested), and the analyst, who is expected to provide the decision support. The aim of this particular process is to establish a shared representation of the client's concern, using the analyst's methodological knowledge, a representation enabling to undertake an action towards the concern.

2.1. Cognitive artefacts

Although decision aiding is a distributed process of cognition, we will present this concept using an operational approach based on the identification of the cognitive artefacts of the process (the outcomes or deliverables) (for more details, the reader is referred to the studies of Bouyssou *et al.*, 2000, and Tsoukiàs, 2007, 2008). The outcomes of this process are as follows:

- a representation of the problem situation: $\mathcal{P}$;
- the establishment of a problem formulation: Γ;
- the construction of an evaluation model: T;
- the establishment of a final recommendation: Φ.

In this paper, we will focus on the establishment of Γ and the construction of T. Our interest in this part of the process is because both these artefacts represent the easier to formalize and more structured outcomes of the process. Therefore, they can be easily modelled using a formal language. Two points should be considered:

- although these two artefacts appear subsequent, they are constructed through continuous interactions;
- the way the DAP is conducted influences the process outcomes.

We can now go through more details as far as these two artefacts are concerned.

2.1.1. Problem formulation (Γ). For a given representation of the problem situation, the analyst might propose to the client one or more 'problem formulations'. This is a crucial point of the DAP. The representation of the problem situation has a descriptive (at the best explicative) purpose. The construction of the problem formulation introduces what we call a model of rationality. A problem formulation reduces the reality of the decision process, within which the client is involved, to a formal and abstract problem. The result is that one or more of the client's concerns are transformed to formal problems on which we can apply a method (already existing, adapted from an existing one or created ad hoc) of the type studied in decision theory. From a formal point of view, a problem formulation is a triplet $\Gamma =< \mathbb{A}, \mathbb{V}, \Pi >$ where

- $\mathbb{A}$ is the set of potential actions the client may undertake within the problem situation as represented in $\mathcal{P}$. It should be noted that these are not 'given' but have to be constructed. A typical situation is the refinement of abstract options to more precise actions. A does not necessarily have a formal structure.
- $\mathbb{V}$ is the set of points of view under which the potential actions are expected to be observed, analysed, evaluated, compared etc., including different scenarios for the future.
- Π is the problem statement, the type of application to perform on the set A, an anticipation of what the client expects.

A problem statement can be operational or not (such as describing or conceiving the elements of $\mathbb{A}$). Operational problem statements are partitioning operations to be applied on the set A within the evaluation model $\mathcal{M}$ (see below). As such they can partition the set A:

- in predefined categories (large-medium-small, illness (A)-illness(B)) or in categories to be inferred comparing the elements of A;
- in ordered categories (accepted-rejected, bad-medium-good) or unordered categories (greens-blacks, monkeys-elephants).

Therefore, Π can be a choice statement (ordered and not predefined categories), a ranking (ordered and not predefined categories), a classification (predefined and not ordered categories), a clustering (no predefined and not ordered categories) etc. (for details see Bana a Costa, 1996; Tsoukiàs, 2007).

2.1.2. Evaluation model ($\mathcal{M}$). The term evaluation model refers to the decision aiding models as they are conceived in operational research, decision theory or AI methods. Classic decision aiding approaches focus on the construction of this model and consider the problem formulation as given. An evaluation model is a tuple $\mathcal{M} =< A, D, H, U, R >$, where

- A is a precise set of alternatives or decision variables on which the model will apply; A has a precise structure: enumeration of actions, domain of real numbers, combinatorial structure etc.;
- D is a set of dimensions (attributes) under which the elements of A are observed, measured, described etc.; a scale is always associated to each element of D;
- H is a set of criteria (if any) under which each element of A is evaluated in order to take into account the client's preferences;
- U is an uncertainty structure;
- R is a set of operators (aggregation functions) such that it is possible to obtain a comprehensive relation and/or function on A, possibly allowing to infer a final recommendation.

We emphasize the different use of terms such as goals (or objectives), attributes and criteria. Goals (objectives) represent 'desired states of the world' and are implicitly or explicitly considered while

constructing the set A (for instance, through the description of alternative plans enabling to achieve a certain task or through the composition of different investment options in order to satisfy a portfolio construction). Attributes and criteria instead represent the fact that achieving a goal or an objective is not only a feasibility problem but also a preferability one. When this is the case, it is necessary to describe the consequences of potential actions along different dimensions (establishing attributes) and to evaluate the client's preferences on some (possibly all) of such consequences (establishing criteria).

2.2. Conducting the process

The DAP is the result of a dialogue between an analyst and a decision maker. During this process, the four artefacts may evolve, change and undergo revisions. Moreover, because a DAP always refers to a decision process that has a time and space extension, it is natural that the outcomes of the DAP remain *defeasible cognitive artefacts* in the sense that new information, beliefs and values may invalidate them and require an update or a revision.

Going back to the model of DAP, we present example 1 that serves the purpose of illustrating possible changes, revisions or updates associated with the formulation problem and its corresponding evaluation model during a DAP.

Example 1. (Bouyssou *et al.*, 2006)
A client looking for decision support within a problem situation described as 'the client's bus company is looking for a bus'. He presents a set of offers received from several suppliers, each offer concerning a precise type of bus. The analyst will establish a problem formulation in which

- $\mathbb{A}$ is the list of offers received;
- $\mathbb{V}$ is the list of point of view that are customary in such cases, let us say cost, quality and transportation capacity;
- Π is a choice problem statement (an offer has to be chosen).

It is possible to construct an evaluation model with such information in which

- A are the feasible offers;
- D are the dimensions on which the offers are analysed: price and management costs, technical features (for the quality point of view) etc.;
- H are the criteria that the client agrees to use in order to represent these preferences;

- there is no uncertainty;
- R can be a multi-attribute value function, assuming that the client is able to establish the marginal value function on each criterion.

When this model is presented to the client, his reaction could be 'in reality we can buy more than one bus and there is no reason that we should buy two identical buses, since these could be used for different purposes such as long range leisure travels or urban school transport'. With such information, it is now possible to establish a new evaluation model in which

- A are all pairs of feasible offers;
- D are the dimension under which the offers are analysed (price, management costs, technical features etc.) but now concerning pairs of offers plus a classification of the buses in categories (luxury liner, mass transit, etc.);
- H are the same as previously plus a criterion about 'fitting the demand' because two different types of buses may fit the demand better;
- now, uncertainty is associated to the different scenarios of bus use;
- R can be multi-attribute utility function, provided that the client is able to establish the marginal value function on each criterion.

The process may continue revising models and problem formulations until the client is satisfied.

Note that it is necessary to update the contents of different models as the DAP involves in time and space. When confronted with a result, the decision maker realized that the model is not exactly what he expected. Therefore, he makes changes or gives new information in order to adapt the model to his needs. The consequence of this update is that the two models should be revised, namely the problem formulation and the evaluation model.

Moreover, the outcome of decision aiding is a recommendation Φ that is submitted to the user. There are three possibilities for this:

Φ_1 the recommendation is validated and implementable
Φ_2 the recommendation is validated but fails to be implemented
Φ_3 the recommendation is not validated

The way the recommendation is submitted to the user is out of the scope of this paper.

The user therefore receives a pair $\langle \Phi_j, T \rangle$ where

- Φ_j represents the state of the recommendation with the user;
- T represent the reasons for which the recommendation is in such a state.

In case the recommendation is in state Φ_1, then the reasons in T are the overall appreciation of the user. Possible reasons for a recommendation in state Φ_2 or Φ_3 can be (i) no feasible solution in A satisfies the user or the recommendation is no more feasible; (ii) the available measures of elements in A are considered irrelevant, erroneous or affected by too large uncertainty; (iii) the preference models applied on A are not reliable and the user feels not to be correctly represented; (iv) uncertainty is misrepresented; or (v) the aggregation procedure is revealed to be meaningless or irrelevant to the user.

In such situations, different questions can be asked: *how to construct such reasons to be meaningful in the decision context considered? How to identify the problem? Or how to provide the consequences to the decision maker of a modification or a changes? Is it possible to challenge the aggregation procedure and how to update it? etc*. Thus, decision aiding is more than simply solving a complex decision model more or less faithful to the decision maker's values and preferences. It involves understanding, interpreting, justifying, explaining, convincing, revising and updating the outcomes of what we call a DAP. Currently, the model of DAP provides a rich theoretical framework in terms of aggregation of preferences and constructing recommendation for various decision problems. However, from a practical point of view, it offers little about how such activities are formally represented. We might be interested to establish a formal representation of all such activities for at least two reasons:

- enhance decision support capabilities of the analyst representing explicitly and accountably the reasons for which he recommend (or not) a solution (if any);
- enhance decision support capabilities of an (semi) automatic device to handle (at least partially) the dialogue with the user.

This work addresses these needs by relying on the concepts and tools of argumentation theory.

3. ARGUMENTATION AND ARTEFACT DEPENDENCIES

We have seen in the previous example that different version of the cognitive artefacts can be established during the DAP. These different versions are because client does not know how to express clearly, at the beginning of the process, what is his or problem and what are his her preferences. So, as the model is constructed, the decision maker may revise and update his preferences and/or objectives. However, such different versions are strongly related to each other because they carry essentially the same information and only a small part of the model has to be revised (Bouyssou *et al.*, 2006; Tsoukiàs, 2007). The problem that arises here is that this revision (or update) must be taken into account by the model. In other words, there is a need for a formal representation of how the evolution occurs between different versions.

In the following, we discuss an argumentation framework that captures the dependencies between the artefacts and illustrate its working by means of an example.

3.1. The argumentation framework

This section gives the basic concepts of the underlying argumentation framework in which an agent represents and reasons with its decision aiding theory. This framework was proposed in the study conducted by Kakas *et al.* (1994) and developed further in that of Kakas and Moraitis (2003), in order to accommodate a dynamic notion of priority over the rules (and hence the arguments) of a given theory.

In this framework, (the components of) an agent theory is layered in three levels. *Object-level decision rules* are defined at the first level. The next two levels describe priority rules on the decision rules of the first level and on themselves thus expressing a preference policy for the overall decision making of the agent. This policy is separated into two levels: level 2 to capture the *default* preference policy under normal circumstances, whereas level 3 is concerned with the *exceptional* part of the policy that applies under specific contexts. The argumentation-based decision making will then be sensitive to context changes.

In general, an argumentation theory is defined as follows.

Definition 3.1

A theory is a pair $(\mathcal{T}, \mathcal{P})$. The sentences in $\mathcal{T}$ are propositional formulae, in the background monotonic logic $(\mathcal{L}, \vdash)$ of the framework, defined as $L \leftarrow L_1, \ldots, L_n$, where $L, L_1, \ldots, L_n$ are positive or explicit negative ground literals. Rules in P are the same as in T apart from the fact that the head L of the rules has the general form $L = h\_p(\text{rule1}, \text{rule2})$, where rule1 and rule2 are ground functional terms that name any two rules in the theory. This higher priority relation given by h_p is required to be irreflexive. The derivability

relation, $\vdash$, of the background logic is given by the single inference rule of modus ponens.

For simplicity, it is assumed that the conditions of any rule in the theory do not refer to the predicate h_p thus avoiding self-reference problems. For any ground atom $h\_p(\text{rule}1, \text{rule}2)$, its negation is denoted by $h\_p(\text{rule}2, \text{rule}1)$ and vice versa.

An *argument* for a literal L in a theory $(\mathcal{T}, \mathcal{P})$ is any subset, T, of this theory that derives L, i.e. $T \vdash L$ under the background logic. The subset of rules in the argument T that belong to $\mathcal{T}$ is called the *object-level* argument. Note that in general, we can separate out a part of the theory $\mathcal{T}_0 \subset \mathcal{T}$ and consider this as a non-defeasible part from which any argument rule can draw information that it might need. We call $\mathcal{T}_0$ the background knowledge base.

The notion of attack between arguments in a theory is based on the possible conflicts between a literal L and its negation and on the priority relation of h_p in the theory.

Definition 3.2
Let $(\mathcal{T}, \mathcal{P})$ be a theory, $T, T' \subseteq \mathcal{T}$ and $P, P' \subseteq \mathcal{P}$. Then (T', P') attacks (T, P) iff there exists a literal L, $T_1 \subseteq T'$, $T_2 \subseteq T$, $P_1 \subseteq P'$ and $P_2 \subseteq P$ such that

(i) $T_1 \cup P_1 \vdash_{\min} L$ and $T_2 \cup P_2 \vdash_{\min} \neg L$
(ii) $(\exists\, r' \in T_1 \cup P_1, r \in T_2 \cup P_2$ such that $T \cup P \vdash h\_p$ $(r, r')) \Rightarrow (\exists\, r' \in T_1 \cup P_1, r \in T_2 \cup P_2$ such that $T' \cup P' \vdash h\_p(r', r))$.

Here $S \vdash_{\min} L$ means that $S \vdash L$ and that no proper subset of S implies L. When L does not refer to h_p, $T \cup P \vdash_{\min} L$ means that $T \vdash_{\min} L$. This definition states that a 'composite' argument (T', P') is a counter-argument to another such argument when it derives a contrary conclusion, L, and $(T' \cup P')$ makes the rules of its counterproof at least 'as strong' as the rules for the proof by the argument that is under attack. Note that the attack can occur on a contrary conclusion $L = h\_p(r, r')$ that refers to the priority between rules.

Definition 3.3
Let $(\mathcal{T}, \mathcal{P})$ be a theory, $T \subseteq \mathcal{T}$ and $P \subseteq \mathcal{P}$. Then (T, P) is admissible iff $(T \cup P)$ is consistent and for any (T', P'), if (T', P') attacks (T, P), then (T, P) attacks (T', P'). Given a ground literal L, then L is a credulous (respectively sceptical) consequence of the theory iff L holds in a (respectively every) maximal (with respect to set inclusion) admissible subset of T.

Hence when we have dynamic priorities, for an object-level argument (from $\mathcal{T}$) to be admissible, it needs to take along with it priority arguments (from $\mathcal{P}$)

to make itself at least 'as strong' as the opposing counter-arguments. This need for priority rules can repeat itself when the initially chosen ones can themselves be attacked by opposing priority rules, and again, we would need to make now the priority rules themselves at least as strong as their opposing ones.

An agent's argumentation theory will be defined as a theory $(\mathcal{T}, \mathcal{P})$, which is further layered in separating $\mathcal{P}$ into two parts as follows.

Definition 3.4
An agent's argumentative policy theory, T, is a theory $T = (\mathcal{T}, (\mathcal{P}_R, \mathcal{P}_C))$ where the rules in $\mathcal{T}$ do not refer to h_p, all the rules in $\mathcal{P}_R$ are priority rules with head $h\_p(r_1, r_2)$ such that $r_1, r_2 \in \mathcal{T}$ and all rules in $\mathcal{P}_C$ are priority rules with head $h\_p(R_1, R_2)$ such that $R_1, R_2 \in \mathcal{P}_R \cup \mathcal{P}_C$.

We therefore have three levels in an agent's theory. In the first level, we have the rules $\mathcal{T}$ that refer directly to the subject domain of the theory at hand. We call these the *object-level decision rules* of the agent. In the other two levels, we have rules that relate to the policy, under which the agent uses its object-level decision rules, associated to normal situations (related to a default context) and specific situations (related to specific or exceptional contexts). We call the rules in $\mathcal{P}_R$ (named R in the following) and $\mathcal{P}_C$ (named C), *default* or *normal context priorities* and *specific context priorities*, respectively.

3.2. Modelling of the decision aiding process
In a nutshell, an automated decision aiding system implements two mappings that correspond to the steps of the DAP. The first mapping, which corresponds to the problem formulation, is one of the form Γ : Problem $\rightarrow < \mathbb{A}, \mathbb{V}, \Pi >$. The second mapping corresponds to the evaluation model construction and is of the form $\mathcal{M} :< A, V, \Pi > \rightarrow < A, D, H, U, R >$.

The first mapping can be implemented by a set of logical rules that associate various parameters, such as the features of the input problem, the situation at hand, the profile of the user etc., with the parameters of the problem formulation. For instance the rule

$$\text{select}(A, A_i) \leftarrow \text{feature}(P, F_1), \ldots, \text{feature}(P, F_n)$$

states that if the problem at hand P has the features F_1, F_n, then the parameter $\mathbb{A}$ of the problem formulation of P is instantiated by the set $\mathbb{A}_i$. In the following, we use the notation $C_{\mathbb{A}, \mathbb{A}_i}(P)$ as a shorthand for the set of conditions that need to be satisfied by problem P in order for the parameter $\mathbb{A}$ to be instantiated by

the set $\mathbb{A}_i$ in the problem formulation of P. Therefore, the aforementioned rule can be represented more compactly by $\text{select}(\mathbb{A}, \mathbb{A}_i) \leftarrow C_{\mathbb{A}, \mathbb{A}_i}(P)$.

Having incomplete information about the world, such a model needs to account for the *lack of information*. To cope with this, in the automated DAP *assumptions* are made about other conditions that may influence the selection of the parameters of the problem formulation. These assumptions, and the mode of reasoning associated with them, can be captured in the argumentation framework that is used by rules of the following form:

$$r_i^{\mathbb{A}} : \text{select}(\mathbb{A}, \mathbb{A}_i) \leftarrow C_{\mathbb{A}, \mathbb{A}_i}(P)$$
$$r_j^{\mathbb{A}} : \text{select}(\mathbb{A}, \mathbb{A}_j) \leftarrow C_{\mathbb{A}, \mathbb{A}_j}(P)$$
$$R_{i,j}^{\mathbb{A}} : h_{-}p\left(r_i^{\mathbb{A}}, r_j^{\mathbb{A}}\right)$$
$$R_{j,i}^{\mathbb{A}} : h_{-}p\left(r_j^{\mathbb{A}}, r_i^{\mathbb{A}}\right) \leftarrow SC_{\mathbb{A}, \{j,i\}}(P)$$
$$C_{j,i}^{\mathbb{A}} : h_{-}p\left(R_{j,i}^{\mathbb{A}}, R_{i,j}^{\mathbb{A}}\right)$$

The aforementioned set of rules says that under the conditions $C_{\mathbb{A}, \mathbb{A}_i}$, $\mathbb{A}_i$ is the *default* parameter selection for $\mathbb{A}$ in the problem formulation. If in addition to $C_{\mathbb{A}, \mathbb{A}_i}$ some special conditions $SC_{\mathbb{A}, \{j,i\}}$ hold for the problem at hand, then $\mathbb{A}_j$ is selected instead. Similar rules can be written for the other parameters of the problem formulation, i.e. for $\mathbb{V}$ and Π. Each set of rules that corresponds to each of the parameters $\mathbb{A}$, $\mathbb{V}$ and Π of Γ is denoted by $T_{\mathbb{A}}$, $T_{\mathbb{V}}$ and T_{Π}, respectively.

The next step in automating the DAP is to provide rules creating the mapping between the selected problem formulation and the possible evaluation models. This can also be done along the lines described above. Consider for instance the description of the relation between $\mathbb{A}$ in the problem formulation and the parameter A of an evaluation model. The rules that describe this mapping are of the form:

$$r_j^A : \text{select}(A, A_j) \leftarrow \text{select}(\mathbb{A}, \mathbb{A}_i), C_{A, A_j}(P)$$
$$r_k^A : \text{select}(A, A_k) \leftarrow \text{select}(\mathbb{A}, \mathbb{A}_i), C_{A, A_k}(P)$$
$$R_{j,k}^A : h_{-}p\left(r_j^A, r_k^A\right)$$
$$R_{k,j}^A : h_{-}p\left(r_k^A, r_j^A\right) \leftarrow SC_{A, \{k,j\}}(P)$$
$$C_{k,j}^A : h_{-}p\left(R_{k,j}^A, R_{j,k}^A\right)$$

Similar rules are added for the other parameters of the evaluation model. The rules for the parameters D, H, U and R that correspond to the rule r_j^A are, respectively,

$$r_j^D : \text{select}(D, D_j) \leftarrow \text{select}(\mathbb{V}, \mathbb{V}_i), C_{D, D_j}(P)$$
$$r_j^H : \text{select}(H, H_j) \leftarrow \text{select}(D, D_i), C_{H, H_j}(P)$$
$$r_j^U : \text{select}(U, U_j) \leftarrow \text{select}(H, H_i), C_{U, U_j}(P)$$
$$r_j^R : \text{select}(R, R_j) \leftarrow \text{select}(\Pi, \Pi_i), C_{R, R_j}(P)$$

Additional rules (of the type R and C), similar to those that have been described for A, are added to the argumentation theory and enforce different selections for the evaluation model parameters, wherever special conditions hold. The set of rules that are associated with the choice of the parameters of M are denoted by T_A, T_D, T_H, T_U and T_R, respectively.

Therefore, the resulting argumentation theory T is the union of the above subtheories, i.e. $T = T_{\mathbb{A}} \cup T_{\mathbb{V}} \cup T_{\Pi} \cup T_A \cup T_D \cup T_H \cup T_U \cup T_R$. At each cycle of the DAP that terminates with a rejection of the recommendations, the reasons for this rejection J are added to T, a new theory $T' = T \cup J$ is constructed, and a new reasoning phase starts, this time with the theory T'. In the following, we present an example that illustrates the method.

3.3. An illustrative example

An agent wishes to plan a dinner for this evening. He has four options. He could dine with his girlfriend, with his best friend, alone or stay at home and order a delivery. The agent prefers dining in a restaurant than staying home and dining with company than dining alone. In the first two cases, the venue is not as important as the company. When dining alone, the standard of the venue is very important. It must be in fact excellent in order to compensate for the lack of company. For dining very late at night, the agent prefers to dine alone, either out or order his favourite pizza for delivery. However, his decision criterion now becomes the time required for service.

The set of actions A relevant to the evening dinner can be represented as follows:

```
a₁   dine_with_girlfriend
a₂   dine_with_best_friend
a₃   dine_alone (X),   X ∈ {r₁,···rₙ}
a₄   order_pizza
```

The `dine_alone`(X) action stands for a set of actions obtained by instantiating variable X with a specific restaurant.

The dining problem can be captured within the decision aiding model described earlier as follows. The agent can choose between two alternative problem formulations. The first is the formulation $\Gamma_1 = \langle \mathbb{A}_1, \mathbb{V}_1, \Pi_1 \rangle$, where $\mathbb{A}_1$ is the set of actions a_1, a_2, a_3, $\mathbb{V}_1$ is pleasure and venue standard and Π_1 is a choice (of the best thing to do this evening) or a

Argument Schemes and Critical Questions for Decision Aiding Process

Wassila OUERDANE [a,1], Nicolas MAUDET [a] and Alexis TSOUKIAS [a]

[a] *LAMSADE, Uni. Paris-Dauphine, Paris 75775*

Abstract. Our ambition in this paper is to begin to specify in argumentative terms (some of) the steps involved in a decision-aiding process. To do that, we make use of the popular notion of argument schemes, and specify the related critical questions. A hierarchical structure of argument schemes allows to decompose the process into several distinct steps—and for each of them the underlying premises are made explicit, which allows in turn to identify how these steps can be dialectically defeated *via* critical questions. This work initiates a systematic study which aims at constituting a significant step forward for forthcoming decision-aiding tools. The kind of system that we foresee and sketch here would allow: (i) to present a recommendation that can be explicitly justified; (ii) to revise any piece of reasoning involved in this process, and be informed of the consequences of such moves; and possibly (iii) to stimulate the client by generating contradictory arguments.

Keywords. Decision aiding, argument schemes, critical questions

Introduction

Decision theory and multiple criteria decision analysis have established the theoretical foundations upon which many decision-support systems have blossomed. However, such systems have focussed more on how a "best solution" should be established, and less on how a decision maker should be convinced about that (for exceptions on that see [9,5]). In addition, the decision-support process is often constructive, in the sense that the client refines its formulation of the problem when confronted to potential solutions. This requires the system to cater for revision: it should be possible, for the client, to refine, or even contradict, a given recommendation. These aspects are usually handled by the decision analyst, but if we are to automate (some part of) the process (as is the case in recommender systems, for instance), it is important to understand more clearly how they can be integrated in a tool.

In AI, a different tradition to decision making had identified these problematic issues. One of the key distinctive ingredient is that many AI-based approaches are prone to represent decision making in terms of "cognitive attitudes" (as exemplified in the famous Belief-Desire-Intention paradigm) [11,12], instead of crude utilities (as already elicited by the analyst). This change of perspective paved the way for more flexible decision-making models: goals may change with circumstances, and understanding these under-

---

[1]Corresponding Author.

lying goals offers the opportunity to propose alternative actions, for example. But then a reasoning machinery has to be proposed to handle these complex notions. Regarding the issues of expressiveness and ability to deal with contradiction that we emphasized here, argumentation seemed a good candidate. Indeed, recently, following some early works [13,8], several models have been put forward in the artificial intelligence community that make use of argumentation techniques [16] to attack decision problems. These approaches have contributed to greatly extend our understanding of the subject, in particular they clarify what makes argumentation about actions crucially different from mere epistemic argumentation (when the object under discussion is a belief).

On the one hand, our contribution is much more modest in its current state than the aforementioned approaches. We will not, in the present paper, base our model on cognitive attitudes and try to represent the underlying motivations and informations of agents. We take, instead, a different perspective which results from the following observation: there exist many decision-support tools that clients understand well, find valuable, and would be reluctant to drop for a completely new tool. Hence the following question: is it possible (and to what extent) to integrate some flavour of argumentation within these tools. On the other hand, having to deal with complex aggregation procedures proposed in these approaches, we will also have to make explicit and discuss some aspects that are often left aside by argumentation-based approaches (although it is known that some aggregation procedures can be captured by an argumentative approach [1]). The main one being that the aggregation procedure itself may be the subject of potential exchange of arguments.

The remainder of this paper is as follows. Section 1 offers a brief reminder on decision aiding theory. In particular we identify the different steps that compose the process, and the nature of the involved objects. Section 2 then presents the different argument schemes that are involved in such processes. The section that follows exploits this representation and defines the critical questions that can be attached to the argument schemes. In Section 4, we present the nature of the resulting dialectical process, pointing out the added-value of this argumentation-based approach when compared to classical multicriteria decision-aiding tools. We conclude by discussing perspectives of this work.

1. Decision Aiding Process

An instance of a decision process is characterized by the participating actors, their concerns, and the resources committed by each actors on each object. We are interested in decision aiding. Intuitively, in decision aiding we also make decisions (what, why and how to model and support). Decision aiding is also a decision process but of a particular nature [9,10]. A decision aiding context implies the existence of at least two distinct actors (the client and the analyst) both playing different roles; at least two objects, the client's concern and the analyst's (economic, scientific or other) interest to contribute; and a set of resources including the client's domain knowledge, the analyst's methodological knowledge, money, time... The ultimate objective of this process is to come up with a consensus between the client and the analyst [19]. Four cognitive artifacts constitute the overall process:

> *Problem situation*— the first deliverable consists in offering a representation of the problem situation for which the client has asked the analyst to intervene;

Problem formulation— given a representation of the problem situation, the analyst may provide the client with one or more problem formulation. The idea is that a problem formulation translates the client's concern, using the decision support language, into a "formal problem";

Evaluation Model—for a given problem formulation, the analyst may construct an evaluation model, that is to organise the available information in such a way that it will be possible to obtain a formal answer to a problem statement. An evaluation model can be viewed as a tuple comprising the set of alternatives on which the model applies (denoted $\mathcal{A}$); the set of dimensions (attributes) under which the elements of $\mathcal{A}$ are observed, described, measured, etc. (denoted $\mathcal{D}$); the set of scales $\mathcal{E}$ associated to each element of $\mathcal{D}$; the set of criteria $\mathcal{H}$ under which each element of $\mathcal{A}$ is evaluated in order to take in account the client's preference; and an aggregation procedure ($\mathcal{R}$). Formally, a criterion is a preference relation, that is a binary relation on $\mathcal{A}$ or a function representing the criterion. (A set of uncertainty structures may also be used. Depending on the language adopted, this set collects all uncertainty distributions or the beliefs expressed by the client. We shall not discuss it further here).

Final recommendation—the evaluation model will provide an output which is still expressed in terms of the decision support language. The final recommendation is the final deliverable which translate the output into the client's language.

The study of this process shows that it suffers from some limits. The first one is the lack of a formal justification or explanation of the final recommendation. Indeed, the process focuses more on how to reach the final decision and fails in some way to provide a justification for the decision-maker. Second, during the decision aiding process several different versions of the cognitive artifacts may be established. These different versions are due to the fact that client doesn't known how to express clearly, at the beginning of the process, what is his problem and what are his preferences. So, as the model is constructed, the decision maker revise and update his preferences and/or objectives. However, such different versions are strongly related to each other since they carry essentially the same information and only a small part of the model has to be revised [19,10]. The problem that arises here is that this revision (or update) is not taken into account by the model. In other words, there is no formal representation of how the evolution occurs between different versions. Finally, the last problem encountered in this process is the incomplete information. More specifically, the process does not support situations or problems decision where some fields of one or more of the different models are not completed.

In this paper we concentrate on the evaluation step. The approach based on argumentation that we sketch in the next few sections is particularly well suited to tackle these aspects: (i) by presenting the reasoning steps under the form of argument schemes, it makes justification possible, and offers the possibility to handle default reasoning with incomplete models; and (ii) by defining the set of attached critical questions, it establishes how the revision procedure can be handled.

2. Argument Schemes

Argument schemes are argument forms that represent inferential structures of arguments used in everyday discourse, and in special contexts like legal argumentation, or scientific

argumentation. *disjunctive syllogism* are very familiar. But some of the most common and interesting argumentation schemes are neither deductive nor inductive, but *defeasible and presumptive* [22].

It is now well established that argument schemes can play two roles: (i) when constructing arguments, they provide a repertory of forms of argument to be considered, and a template prompting for the pieces that are needed; (ii) when attacking, arguments provide a set of critical question that can identify potential weaknesses in the opponents case. Then, as Walton puts it, " we have two devices, *schemes* and *critical questions*, which work together. The first device is used to identify the premises and conclusion. The second one is used to evaluate the argument by probing into its potentially weak points" [22]. The set of critical questions have to be answered, when assessing whether their application in a specific case is warranted. Prakken and Bench-capon [6] specify that argument schemes are not classified according to their logical form but according to their content. Some argument schemes express epistemological principles or principles of practical reasoning: different domains may have different sets of such principles. Our aim in this paper to identify those schemes that are involved in multicriteria decision-aiding processes.

We need different classes of argument schemes to construct the whole evaluation model. Argument schemes can very broadly be distinguished depending on (i) whether they aggregate several criteria, or are concerned with a single criteria (multicriteria vs. unicriteria); (ii) whether they follow a pairwise comparison principle or whether they use an intrinsic evaluation, the action being compared to a separation profile (intrinsic vs. pairwise); and (iii) whether they are concerned with the evaluation of the action or its mere acceptability (evaluation vs. acceptability). In theory, all combinations seem possible, even though some are much more natural than others.

In this paper, we shall focus our attention on the following schemes:

> argument schemes for *Unicriteria Pairwise Evaluation* (UC-PW-EV), which establishes that an objet is at least prefered to another object from the single viewpoint of the considered criteria (note that there may be an intrinsic version of this scheme, for instance for classification, but also to cater for all the argumentation-based aggregation techniques);
> argument schemes for *Unicriteria Intrinsic Acceptability* (UC-IN-AC), which establishes that the action can be considered in the evaluation process (here also, it may be possible to have a similar scheme for relative "pairwise" acceptability);
> argument scheme for *Multicriteria Pairwise Evaluation* (MC-PW-EV), which basically concludes that an object is at least as good as another object on the basis of several criteria taken together. It is constituted of two sub-argument schemes:
>
>> argument schemes for *Positive Reasons Aggregation Process* (PR-AP), which concludes that there are enough positive reasons to support the claim of MC-PW-EV, and that can be of many types depending on the aggregation technique used (ex. simple majority, weighted sum, and so on);
>> argument schemes for *Negative Reasons Aggregation Process* (NR-AP) which concludes that the negative reasons should block the conclusion of MC-PW-EV (again, this really constitute a family of argument scheme);

argument schemes for *Global Recommendation* (GR) which provides the output of the process (of different type depending on the decision problem considered). We shall not discuss this level in this paper.

In the rest of this paper we limit the discussion to the case involving only two actions. This is a basic building block that will be required if we are to construct more general decision-aiding.

Now we turn our attention to argument schemes. In fact, as must be clear from the discussion above, there is an underlying hierarchical structure that ties the different argument schemes. In short, we can distinguish three levels of argument schemes that will be embedded. At the highest level the multicriteria pairwise evaluation, which is based on the aggregation of positive and negative reasons, which in turn is based on unicriteria evaluation of actions versus other actions (or special profiles).

2.1. Argument Schemes for Unicriteria Action Evaluation

The first way to perform an action evaluation is to compare two actions from the point of view of the chosen criterion: this is modeled by the scheme for Unicriteria Pairwise Evaluation (UC-PW-EV), see Tab. 1. This argument scheme is the basic piece of reasoning that is required in our decision-aiding context. It concludes that an action a is at least as good as an action b from the point of view of a given criterion h_i, based on some preference relation $\succeq_i$ [17].

Premises	a criteria	h_i
	an action	a
	whose performance is	$g_i.a/$
	an action	b
	whose performance is	$g_i.b/$
	a preference relation	$\succeq_i$
Conclusion	a is at least as good as b	$a \succeq_i b$

Table 1. Scheme for Unicriteria Pairwise evaluation (UC-PW-EV)

When an action needs to be intrinsically evaluated, there is a need to define the categories and *separation profiles*. Such a separation profile defines on each criterion a sort of neutral point: this is by not necessarily an existing action, but it allows to define to which category to affect the action. A particular case is when we only consider "pro" and "con" categories. The scheme for Unicriteria Intrinsic Action Evaluation, as given in Tab. 2, details such a scheme.

Premises	an action	a
	whose performance is	$g_i.a/$
	along a criteria	h_i
	a separation profile	p
	whose performance is	$g_i.p/$
	a preference relation	$\succeq_i$
Conclusion	a is acceptable according to h_i	$a \succeq_i p$

Table 2. Scheme for Unicriteria Intrinsic Action Evaluation (UC-IN-EV)

2.2. Argument Schemes for Acceptability

The case of action acceptability is very similar to that of action evaluation: it can also be performed intrinsically or in pairwise manner. We start with the *Argument Scheme for Intrinsic Acceptability* (UC-IN-AC). The scheme is very similar to that of Unicriteria Intrinsic Evaluation. In fact, in this case the separation profile can play the role of a *veto threshold*: when the action does not reach that point, there are good reasons to exceptionally block the claim (disregarding the performance of the action on other criterion). For the sake of readability, we shall not repeat this very similar scheme here. A different kind of acceptability relies instead on the relative comparison of actions: it may be the case that an action is considered to be inacceptable because the difference in performance is so huge with another action. In this case, we talk about an *Argument Scheme for Pairwise Acceptability* (UC-PW-AC). We believe this is self-explanatory given the examples provided so far, and shall not give any further detail here.

2.3. Arguments Scheme for Aggregating Positive Reasons

At this level the piece of reasoning involved must make clear how we can conclude that enough positive reasons are provided. Perhaps the most obvious such scheme, at least one that is ubiquous in multicriteria making is the *principle of majority*. It only says that a is at least as good as b when there is a majority of criterion supporting this claim. Table 3 gives the detail of the corresponding argument scheme.

Premises	a set of criteria considered to be of equal importance	$\{h_1; h_2; \ldots; h_n\}$
	a set of pairwise evaluation of actions a and b	
	the majority support the claim	
Conclusion	there are good reasons to support a is at least as good as b	$a \quad b$

Table 3. Scheme for Argument from the Majority Principle (PR-AG (maj))

Note that this scheme makes explicit that criteria are considered to be of equal importance. This is not necessarily the case, and more generally many other aggregation techniques may be used to instantiate $\mathcal{R}_P$. These other schemes will potentially require additional information, which justifies that we have many different scheme and not a single generic one. For instance, a possible scheme would conclude that a is at least as good as b when it is at least as good on (some of) the most important criteria (*argument from sufficient coalition of criteria*).

Here we only present a different one to illustrate the variety of argument schemes that may be used. This simple typical example is the lexicographic method that we detail below. The method works as follow: look at the first criterion, if a is strictly better than b on this, then a is declared globally preferred to b without even considering the following criteria. But if a and b are indifferent on the first criterion, you look at the second one, and so on.

Note that the basic input information that needs to be provided to these schemes is that of a pairwise comparison on a single criterion dimension (the output of UC-PW-EV). Indeed, this will be in most case the basic building block upon which the recommendation can be build. There is however a different type of scheme that would aggregate instead intrinsic valuations of both actions: that would be the case of argument-based

Premises	a set of criteria	$\{h_1; h_2; ::::; h_n\}$
	a linear order on the set of criteria	$h_1 > h_2 > ::: > h_n$
	a set of pairwise evaluation of actions a and b	
	a is strictly better than b on h_i	$a \succ_i b$
	a is indifferent to b on h_j for any $j < i$	$a \sim_j b$ when $j < i$
Conclusion	there are good reasons to support a is at least as good as b	$a \succeq b$

Table 4. Scheme for Argument from the lexicographic method (PR-AG (lex))

aggregation procedures that take as input sets of arguments "pro" and "con". Clearly, the basic argument scheme required will be different here, for it needs to provide an intrinsic evaluation of the action.

2.4. Argument Scheme for Multi-Criteria Pairwise Evaluation

The argument scheme that lies at the top of our hierarchy is inspired by outranking multi-criteria techniques [10], and indeed its argumentative flavour is obvious. The claim holds when enough supportive reasons can be provided, and when no exceptionally strong negative reason is known. This already suggests that there will be (at least) two ways to attack this argument: either on the basis on a lack of positive support, or on the basis of the presence of strong negative reasons (for instance, a "veto"). Typically, supportive reasons are provided by action evaluation, and negative reasons are provided by action (lack of) acceptability. We shall discuss this further when we turn our attention to critical questions.

Premises	an action	a
	an action	b
	a set of criteria	$\{h_1; h_2; ::::; h_n\}$
	there are enough supportive reasons according to	$\mathcal{R}_P$
	there are no sufficiently strong reasons to oppose it	$\mathcal{R}_N$
Conclusion	a is at least as good as b	$a \succeq b$

Table 5. Scheme for pairwise evaluation multicriteria (MC-PW-EV)

Here, $\mathcal{R}_P$ stands for the aggregation process that should be used to aggregate the (positive) reasons supporting the claim, whereas $\mathcal{R}_N$ stand for the aggregation process concerned with the aggregation of *exceptionally* negative reasons (vetos). The conclusion of the scheme expresses that a is at least as good as b according to the preference relation $\succeq_{MCPWEV}$ induced by the scheme.

3. Critical Questions

Along with each different argument schemes comes a set of *critical questions* [22,21]. These questions as we said before, allow us to identify potential weaknesses in the scheme. Below we present the set of critical questions attached to the schemes MC-PW-EV, PR-AG (maj), and UC-PW-EV. We note that different types of critical questions can be identified [14], depending on whether they refer to standard assumptions of the scheme or to exceptional circumstances. This has in particular a significant difference on how the burden of proof is allocated. We now list some of the questions that can be attached to the different premises.

Argument Scheme for Multi-Criteria Pairwise Evaluation. In this context the different type of questions is clear. The burden of proof lies on the proponent when it must provide supportive evidence (positive reasons) for the main claim. On the other hand, the opponent should be the one providing negative reasons to block the conclusion.

1. *actions* (assumption): is the action possible?
2. *list of criteria* (assumption): (i) Is this criteria relevant?, (ii) Should we introduce a new criteria?, (iii) Are these two criteria are in fact the same?
3. *positive reasons* (assumption): (i) Are there enough positive reasons to support the claim? (ii) Is the aggregation technique relevant ?
4. *negative reasons* (exception): Are there not enough reasons to block the claim? Is the aggregation technique relevant?

Note also that while the use of a specific aggregation technique may be challenged at this level ("why are we using a majority principle here?"), the actual exchange of argument regarding this aspect will involve the sub-argument scheme concerned with this aggregation. We now turn our attention to the critical questions that may then be used.

Together with the *Scheme for Argument from the Majority Principle.* come two obvious questions are:

1. *list of criteria* (exception): Are the criteria of equal importance?
2. *majority aggregation* (exception): Is the simple majority threshold relevant for the current decision problem?

As for the *Argument Scheme for Unicriteria Pairwise Action Evaluation*, we can propose this tentative set of questions :

1. *actions* (assumption): Is the action possible?
2. *criterion* (assumption): Is the criteria relevant?
3. *action's performance* (assumption): Is the performance correct?
4. *preference relation* (assumption): Is the preference relation appropriate?

It should be noted that a negative answer to some of these questions leads to a conflict whose resolution requires sometimes the transition to a different stage of the negotiation process. For instance, when you challenge whether the action is possible to start with, you are dealing with the problem formulation (cf. section 1), where the set of alternatives is defined. It is out of the scope of this paper to discuss this problem. We will just mention that through the different critical questions, we have the opportunity to review and correct not only the evaluation model, but also other stages of the process.

4. The Dialectical Process

In this section we give a glimpse of the dialectical process that will exploit the argument schemes and critical questions that we have put forward so far. It is based on the popular model of dialogue games, and more precisely it is based on recent extensions that incorporate argument schemes within such models [18]. The full specification of the dialogue game is the subject of ongoing work. The process initiates with the client specifying the

basic elements of the evaluation model[2] (see Sect. 1): it specifies a set of actions (in the context of this paper we limit ourselves to two actions though), a set of criteria, and the aggregation operators that shall be used. Contrary to classical decision tools, these sets will only be considered to be the *current* evaluation model, and it is taken for granted that it can be revised throughout the process. Now, as we see it, an argumentation-based decision-aiding process should:

1. *justify* its recommendation. Crucially, by presenting its justifications in the form of arguments, the system will make it possible for the user to pinpoint those steps that pose problems. The system builds up the current recommendation by embedding argument schemes of the three levels. The argument schemes are build on the basis of the information provided by the user, and in some cases by using default instantiation (when the scheme allows for it). If challenged by the user, the system provides the different steps of reasoning by revealing parsimoniously the lower level schemes that compose the recommendation. Each time a scheme is presented, the entire set of critical is at the disposal of the user to challenge the current conclusion. There are very different reasons to revise in such a process: in some case the user may simply want to correct/refine one of its previous statement, introduce new information. In other cases it will contradict one of the system's assumption;

2. *inform* the user of the consequences of its modifications. In many cases the user would not foresee the various consequences of a seemingly local modification: in these cases, the system helps the user by making explicit the hidden and critical consequences of its move;

3. potentially *propose* automatically generated arguments, in order to stimulate users (a technique that has proven very useful, for instance in the context of recommender systems [20]). The system holds a collection of argument schemes that can be readily used to generate counter-arguments that would stimulate the user in its decision-making process. In particular, the system may seek to generate counter-arguments based on the use of different argument schemes that would contradict the current recommendation, by relaxing some assumptions or informations provided by the user. The idea being here to emphasize for the user what are the assumptions in the evaluation model currently used.

The different cases mentioned above are illustrated by means of an example.
User: [specify its evaluation model: two actions a and b, five criteria c_1; c_2; c_3; c_4; c_5, among which c_1 is considered to be of primary importance, followed by c_2, and so on. Acceptability is initially intrinsic: no action should have a score on a criterion lower than 1. The performance table is as follows:]

	c_1	c_2	c_3	c_4	c_5
a	7.5	6	2	3	5
b	7	4	8	4	7

1. System: Given your informations, a is at least as good as b. [Recommendation]
2. User: Why? [Challenge]

---

[2]Of course a more ambitious dialectical system would have to consider the previous steps as well. This is beyond the scope of this paper.

3. System: The most important criteria according to you defend this claim, so by comparing actions on the basis of criteria of decreasing importance, a should be prefered to b [Justified Recommendation]
(although the majority of arguments defend the opposite claim [Gen. counter-arguments])
4. User: I still think that c_1 is the most important criteria so the majority makes little sense. [Reject counter-argument]
however, now I see that the difference between a and b is not significant enough...
5. System: My recommendation remains the same because a is prefered to b on the second most important criterion
6. User: I agree but still a cannot be acceptable.
7. System: Why? [Burden of proof shift]
8. User: b is way better than a on c_3, I could not justify this choice... [Argument Scheme for Relative acceptability]
9. System: Ok. [Revise the recommendation]
But beware that applying this scheme for acceptability may have the following consequences... [Inform of consequences of this revision]

Let us briefly analyze this dialogue. Turn 1 provides the recommendation, which is challenged by the user on turn 2. Not being more explicit, the challenge can be assumed to ask the system to provide more explicit information regarding positive reasons supporting the claim. The system, on turn 3, explains that the claim is based on the use of the PR-AG(lex) scheme. Note that it also generates a possible counter-argument by relaxing some of the information provided by the user (here the fact that criteria have different importance). Observe that this is an indirect way for the system to use a critical question...) The user rejects this counter-argument on turn 4 (by re-affirming the fact that criteria have unequal importance), but attacks the basic UC-PW-EV argument upon which the recommendation is based. The critical question used here is that of the relevance of the preference relation. The system accepts the move (and modifies the user's information by specifying that actions should exhibit at least half a point of difference, otherwise they should be considered as indifferent). But the system restates that the recommendation remains unchanged: this is due to the fact on the second most important criterion, *a* is again better than *b*. (The attack is *unrelevant* in Prakken's sense). The user accepts this but now attacks on the ground of negative reasons, and explains that *a* can not be accepted on the basis of pairwise acceptability (UC-PW-AC). Finally, the system revises its recommendation but may at the same time make explicit the consequences of the proposed change.

5. Related work

One of the most convincing proposal recently put forward to account for argument-based decision-making is the one by Atkinson *et al.* [3,2]. They propose an extension of the "sufficient condition" argument scheme for practical reasoning [21], by distinguishing the goal into three elements: state, goal and value. This scheme serves as a basis for the construction of a protocol for a dialogue game, called Action Persuasion Protocol (PARMA) [4]. The authors show how their proposal can be made computational within the framework of agents based on the BDI model, and illustrate this proposal with an

example debate within a multi-agent system. Prakken et al. [7] offer a logical formalisation of Atkinson's account within a logic for defeasible argumentation. They address the problem of practical syllogism by trying to answer questions such as: how can an action be justified? In particular, the aim is to take into account the abductive nature of the practical reasoning and the side effects of an action. A key element in this formalisation is the use of accrual mechanism for argument to deal with side effects (positive and negative effects).

The first approach attempting to introduce argumentation in the decision aiding process as a whole is the one of Moraitis et *al.* in [15]. The idea is to describe the outcomes of the decision aiding process through an operational model and to use argumentation in order to take into account the defeasible character of the outcomes. The authors tryu to provide a way allowing the revision and the update of the cognitive artifacts of the Decision Aiding Process.

In addition to these works, many other proposals have been put forward in the literature to use argumentation in a decision context, see [16] for a recent survey. From the point of *decision aiding* though, a couple of elements remain largely unexplored. Under that perspective, current argumentation models are not fully satisfying because for instance: (i) most of the approaches assume a decision problem where the aim is to select the "best" action for a given purpose, when in fact a variety of decision problems can be addressed (choice, ranking, sorting,...); and (ii) most models currently proposed in the literature rely on an underlying *intrinsic evaluation* (actions are evaluated against some absolute scale), whereas most decision aggregation procedure make use of *pairwise evaluation* techniques (actions are compared against each others).

6. Conclusion and Future Work

The purpose of this paper was to provide a first approach to represent the steps of a multi-criteria decision aiding process by means of argument schemes and critical questions. We focused here on the evaluation model, and considered the restricting but basic case of the comparison of two actions. To represent the decision evaluation process, we identified a hierarchical structure of argument schemes. Each level refers to one step in the classical multicriteria evaluation. The highest level represents the pairwise evaluation, which is based on the aggregation level, which is in turn based on unicriteria evaluation (pairwise or intrinsic). To these schemes we associated a set of critical questions. One reviewer of this paper raised the following issue: does it make sense in the first place to consider argument schemes that cover the aggregation level? One of the main claim of this paper is that it does, precisely because the way basic argument schemes are collected and aggregated may also be disputed, and be based on assumptions that can be challenged and/or revised. The aim is (as usual with argument schemes and critical questions, as proposed here) to allow us to check the acceptability of each scheme by probing into its potentially weak points, and this from different point of views. We also give the very basic ingredients of the dialectical system currently under development. Future work should extend the model to take into account, in one hand a large set of alternatives, on other hand to handle different decision problems (ranking, sorting,...), in order to build a dialectical system-based decision aiding system for the whole process.

References

[1] L. Amgoud, J.-F. Bonnefon, and H. Prade. An Argumentation-based Approach to Multiple Criteria Decision . In *Proc. of the 8th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 269–280. LNCS, 2005.

[2] K. Atkinson. Value-based argumentation for democratic support. In *Proc. of the 1st International Conf. on Computational Models of Natural Argument*, pages 47–58. IOS Press, 2006.

[3] K. Atkinson, T. J. M. Bench-Capon, and S. Modgil. Argumentation for decision support. In *Proc. of the 17th International Conf. on Database and Expert Systems Applications*, pages 822–831, 2006.

[4] K. Atkinson, T.J.M. Bench-Capon, and P. McBurney. Computational representation of practical argument. *Knowledge, Rationality and Action*, 152(2):157–206, 2006.

[5] V. Belton and T. Stewart. *Muliple Criteria Decision Analysis: An Integrated Approach*. Kluwer Academic, Dordrecht, 2002.

[6] T.J.M Bench-Capon and H. Prakken. Argumentation. In A.R. Lodder & A. Oskamp, editor, *Information Technology & Lawyers : Advanced technology in the legal domain from challenges to darlyroutine*, pages 61–80. Springer Verlag, Berlin, 2005.

[7] T.J.M. Bench-Capon and H. Prakken. Justifying actions by accruing arguments. In P.E Dunne and T.J.M. Bench-Capon, editors, *Pro. of the 1 st International Conf. on Computational Models of Natural Argument(COMMA'06)*, volume 144, pages 311–322, Amesterdam, The Netherlands, 2006. IOS Press.

[8] B. Bonet and H. Geffner. Arguing for Decisions: A Qualitative Model of Decision Making. In *Proc. of the 12th Conference on Uncertainty in Artificial Intelligence (UAI'96)*, pages 98–105, 1996.

[9] D. Bouyssou, T. Marchant, M. Pirlot, P. Perny, A. Tsoukiàs, and Ph. Vincke. *Evaluation and decision models: a critical perspective*. Kluwer Academic, Dordrecht, 2000.

[10] D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs, and P. Vincke. *Evaluation and decision models with multiple criteria: Stepping stones for the analyst*, volume 86 of *International Series in Operations Research and Management Science*. Springer, Boston, 2006.

[11] M. Dastani, J. Hulstijn, and L. van der Torre. How to decide what to do? *European Journal of Operations Research*, 160(3):762–784, 2005.

[12] J. Doyle and R. Thomason. Background to qualitative decision theory. *AI magazine*, 20(2):55–68, 1999.

[13] J. Fox and S. Parsons. On Using Arguments for Reasoning about Actions and Values. In *Proc. of the AAAI Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning*, pages 55–63. AAAI Press, 1997.

[14] T. Gordon, H. Prakken, and D. Walton. The carneades model of argument and burden of proof. *Artificial Intelligence*, 171:875–896, 2007.

[15] P. Moraitis and A. Tsoukiàs. Decision aiding and argumentation. *Proc. of the 1st European Workshop on Multi-Agent Systems*, 2003.

[16] W. Ouerdane, N. Maudet, and A. Tsoukiàs. Arguing over actions that involve multiple criteria: A critical review. In *Proc. of the 9th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 308–319, 2007.

[17] M. Oztürk, A. Tsoukiàs, and Ph. Vincke. Preference modelling. In J. Figueira, S. Greco, and M. Ehrgott, editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 27–72. Springer Verlag, Boston, Dordrecht, London, 2005.

[18] C. Reed and D. Walton. Argument schemes in dialogue. In H.V. Hansen, C.W. Tindale, R.H. Johnson, and J.A. Blair, editors, *Dissensus and the search for common ground*, 2007.

[19] A. Tsoukiàs. On the concept of decision aiding process. *Annals of Operations Research*, 154(1):3–27, 2007.

[20] P. Viappiani, B. Faltings, and P. Pu. Preference-based search using example-critiquing with suggestions. *Journal of Artificial Intelligence Research*, 171:465–503, 2006.

[21] D.N. Walton. *Argumentation schemes for Presumptive Reasoning*. N. J.,Erlbaum, 1996.

[22] D.N. Walton and C.A. Reed. Argumentation schemes and defeasible inferences. In Giuseppe Carenini, Florina Grasso, and Chris Reed, editors, *Workshop on Computational Models of Natural Argument*, 2002.