From Preference Elicitation to Explaining Decisions: a Dialectical Perspective

Wassila Ouerdane

université PARIS-SACLAY

Habilitation à Diriger des Recherches

.

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 v

List of Tables vii

1 Introduction 1
- 1.1 Context and Motivations . . . 1
- 1.2 Research Questions and Contributions . . . 4
  - 1.2.1 Modeling and generating explanations for recommendations for complex decision problems. . . . 5
  - 1.2.2 Modeling the interaction for constructing adaptive decision support systems. . . . 7
- 1.3 Structure and Content of the Document . . . 11

2 MCDA: Concepts and Definitions 13
- 2.1 Multiple Criteria Decision Aiding . . . 13
- 2.2 Preference Learning and Elicitation Process . . . 15
  - 2.2.1 A brief description . . . 15
  - 2.2.2 The aggregation model . . . 17
  - 2.2.3 How to specify an aggregation model? . . . 17
- 2.3 Focus on Some Aggregation Models . . . 19
  - 2.3.1 Additive utility model . . . 19
  - 2.3.2 Non-Compensatory Sorting model . . . 19
- 2.4 Summary . . . 25

3 Efficient Tools for Preference Learning and Elicitation 27
- 3.1 Introduction . . . 27
- 3.2 Learning NCS Model Parameters . . . 28
- 3.3 SAT/MaxSAT Formulations for Inv-NCS . . . 29
  - 3.3.1 SAT-based formulations for Inv-NCS . . . 30
  - 3.3.2 MaxSAT relaxations for Inv-NCS . . . 34
  - 3.3.3 SAT/MaxSAT for Inv-NCS: main experimental insights . . . 35
- 3.4 Learning NCS Model Parameters: new perspectives . . . 37
  - 3.4.1 Learning MR-Sort models with latent criteria direction . . . 39
  - 3.4.2 Learning MR-Sort models with single-peaked preferences . . . 41
- 3.5 Summary . . . 45

4 Supporting Decisions: a Panel of Explainability Tools **47**
4.1 Explainable Artificial Intelligence: Positioning 47
4.2 Explaining Recommendations Stemming from MCDA Models 50
4.2.1 Explaining a recommended choice 51
4.2.2 Explaining pairwise comparisons 58
4.2.3 Explaining an assignment 65
4.3 Summary 73

5 Interactive Recommendations and Explanations **75**
5.1 Dialectical Tools for Decision Aiding 75
5.1.1 Conducting the interaction though a dialogue game. 78
5.1.2 Managing various preference models. 79
5.1.3 Allowing critics/feedback through Critical Questions. 81
5.1.4 Next steps 82
5.2 Explanation Schemes: Generation and Evaluation 83
5.2.1 New explanation schemes/patterns 83
5.2.2 Expressing and presenting an explanation? 85
5.2.3 Evaluating and Assessing explanations 87
5.3 Interactive Explanations 88
5.3.1 Mixed-initiative interaction 88
5.3.2 Modeling and managing inconsistency 89
5.3.3 New perspectives for preference learning and elicitation 90
5.3.4 Interaction: validation and evaluation 91
5.4 Towards Decision Aiding for Collective Decision 92
5.5 Summary 93

Bibliography **95**

Appendices **111**

A Curriculum Vitae **113**

B Publications **125**

List of Figures

2.1 The elicitation process. 15
2.2 Aggregation procedures. 16
2.3 Representation of performances w.r.t. category limits. 22
2.4 Sufficient (green) and insufficient (red) coalitions of criteria 23

3.1 Approaches for comparing learning algorithms 36

4.1 Partial preferences $\succ_1, \succ_2, \succ_3, \succ_4$ over the criteria 1,2,3,4. 55
4.2 Relationships between argument schemes 61
4.3 Covering scheme: a visual representation of Ex. 4.12 64
4.4 Covering scheme: a narrative representation of Ex. 4.12 64

5.1 Dialectical vision for MCDA 75
5.2 Successive speech acts at each iteration 78
5.3 Structure $\mathcal{Q}$ with three properties 80

List of Tables

1.1 Performance table . . . 3
1.2 Our Contributions to the Explainability Topic for MCDA . . . 6
1.3 Our Contributions to preference Learning & Elicitation Topic . . . 9
1.4 Our Contributions to the Interaction Topic . . . 10

2.1 A performance table for car model evaluation . . . 14
2.2 Performance table for models of cars. . . . 22
2.3 Limiting profiles. . . . 22
2.4 Categorization of performances. . . . 23
2.5 Alternative assignments. . . . 23

3.1 Contributions to preference learning and elicitation . . . 28

4.1 Our contributions for explainable MCDA . . . 51
4.2 Structural properties of the reasoning schemes. . . . 61

5.1 Our contributions to adaptive interaction . . . 77

B.1 Publications Summary . . . 133

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) $\rightsquigarrow$ One shall bring proof (complete explanation)
- Explanation shall be understandable $\rightsquigarrow$ 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 $\rightsquigarrow$ 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 $\rightsquigarrow$ 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 **??**. 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) .

Chapter 2

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 $\mathcal{N} = \{1, 2, \dots, n\}$ *points of view* (criteria) evaluating a set of *alternatives* $\mathbb{X} = \{x, y, z, \dots\}$.

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 $\mathcal{N} = \{1, 2, 3, 4, 5\}$ 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 2.1 (Example of preference profiles)

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

Example 2.2 (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.

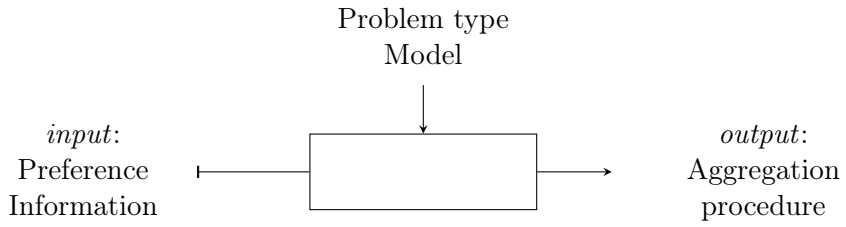


Figure 2.1: The elicitation process.

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.

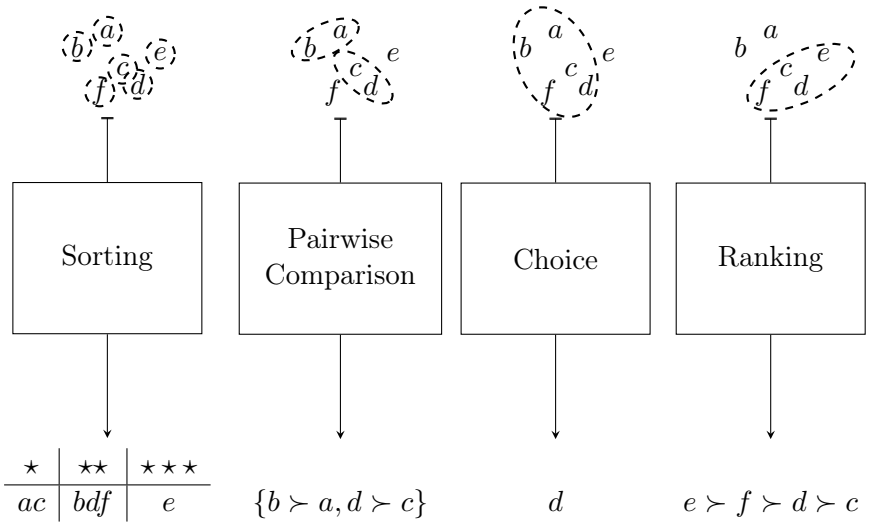


Figure 2.2: Aggregation procedures.

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:

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

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

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 $(\mathbb{X}_i, \succsim_i)$ 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 $\forall i \in \mathcal{N}$, $\mathcal{A}_i^2 = \mathbb{X}$, $\mathcal{A}_i^{p+1} = \emptyset$.

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.

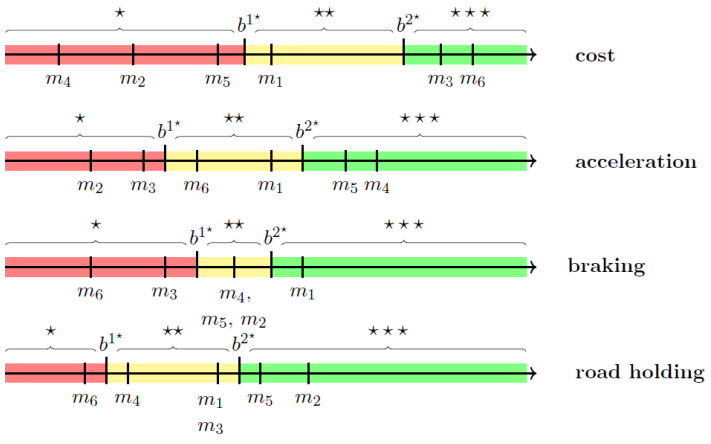


Figure 2.3: Representation of performances w.r.t. category limits.

model	cost	acceleration	braking	road holding
m_1	$\star\star$	$\star\star$	$\star\star\star$	$\star\star$
m_2	$\star$	$\star$	$\star\star$	$\star\star\star$
m_3	$\star\star\star$	$\star$	$\star$	$\star\star$
m_4	$\star$	$\star\star\star$	$\star\star$	$\star\star$
m_5	$\star$	$\star\star\star$	$\star\star$	$\star\star\star$
m_6	$\star\star\star$	$\star\star$	$\star$	$\star$

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	$\star\star$	$\star$	$\star\star$	$\star\star$	$\star\star\star$	$\star$

Table 2.5: Alternative assignments.

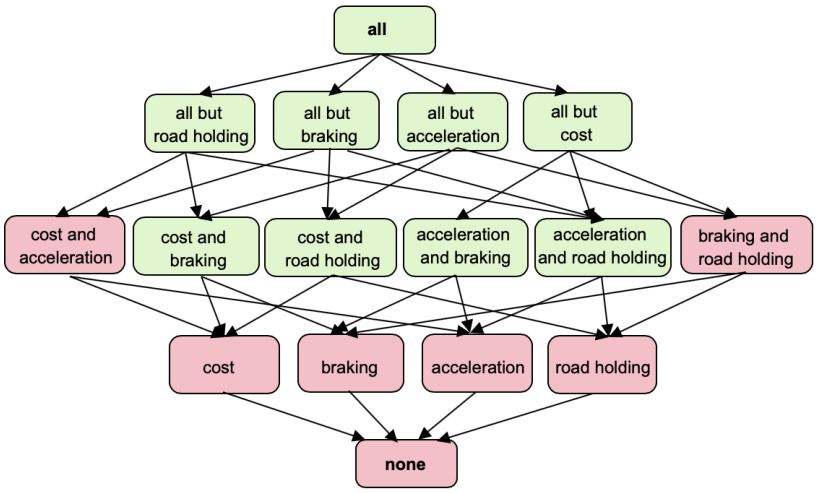


Figure 2.4: Sufficient (green) and insufficient (red) coalitions of criteria

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:

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

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.

Chapter 3

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) [Roland, 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 $(\mathbb{X}^\star, \succsim_i)$: 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 $\alpha : \mathbb{X}^\star \to \{ \textsc{Good} , \textsc{Bad} \}$, 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

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

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

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:

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

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

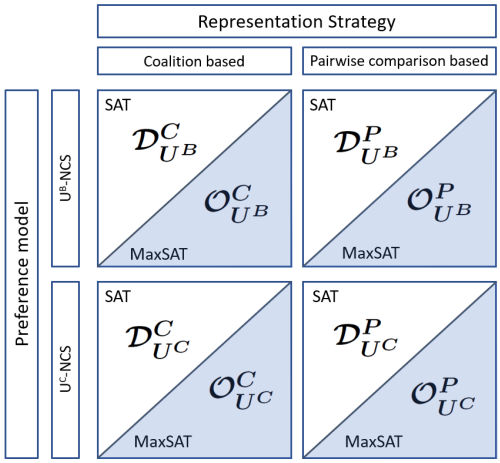


Figure 3.1: Approaches for comparing learning algorithms

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 $\sim$100 examples, we can consider two types of applications: an application that involves a large number of criteria ($|\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 direction

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 $[\underline{b}_i, \overline{b}_i]$).

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.

Chapter 4

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 $\langle$premises, conclusion$\rangle$, 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 **??**).

[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].

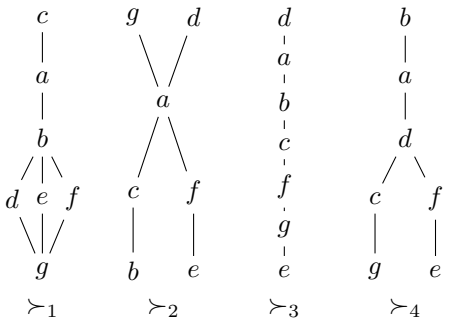
Considering this incompleteness of PI, we investigated the question of searching and defining a simple explanation for a recommended choice. Thus, we are looking to justify that a given option is a weighted Condorcet winner (WCW), i.e. this option is necessarily preferred to each other option, whatever 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.



1 3	$\succ$	2 4
2 3	$\succ$	1
4	$\succ$	2 3

$w_1 + w_3$	$>$	$w_2 + w_4$
$w_2 + w_3$	$>$	w_1
w_4	$>$	$w_2 + w_3$

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 $\texttt{cl}$ such that $\texttt{cl}(V)$ 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) $[14 \succ 23] \in \text{cl}(V)$ follows from $[1 \succ 23]$, by monotonicity.

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

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

Moreover, by examining the elements belonging to $\texttt{cl}(V)$, 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) $\texttt{cl}_0(V)$ contained directly in the PI (no underlying complexity for the user, e.g. $[23 \succ 1]$ in Ex. 4.5), (ii) $\texttt{cl}_1(V)$ that can be deduced from V only using monotonicity (e.g. $[4 \succ 3]$), (iii) $\texttt{cl}_2(V)$ that can be deduced from V only using summation and monotonicity conditions (e.g. $[4 \succ 1]$), and (iv) $\texttt{cl}_3(V)$ that are in $\texttt{cl}(V)$ (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 ($\texttt{cl}_j(V)$, $j \in \{1, 2, 3\}$), identifying all the preference statements ($S' \subset S$) that justify the WCW such that $\mathcal{V}(S') \subset \texttt{cl}(V)$, 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 ($\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}, \texttt{f}, \texttt{g}$), 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 ($s_1 = (\text{✗}, \text{✗}, \text{✓}, \text{✓}, \text{✓}, \text{✓}, \text{✓})$) or more succinctly by the subset of criteria on which it is evaluated positively ($s_1 = \{\texttt{cdefg}\}$). 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 $= \{\mathtt{c}, \mathtt{e}, \mathtt{f}, \mathtt{g}\}$, CON $= \{\mathtt{a}\}$, while NEU $= \{\mathtt{b}, \mathtt{d}\}$

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-

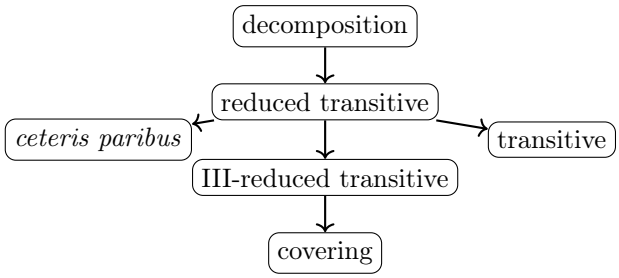


Figure 4.2: Relationships between argument schemes

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).

Scheme	Properties	Requirements for 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 $\omega_{\mathtt{e}}$ and $\omega_{\mathtt{c}}$ on both sides (this is actually an instance of *second order cancellation*, because it is performed across two comparative statements), we obtain $\omega_{\mathtt{f}} + \omega_{\mathtt{g}} + \omega_{\mathtt{b}} \geq \omega_{\mathtt{d}} + \omega_{\mathtt{a}}$, hence $\mathtt{bfg} \succ_{\omega} \mathtt{ad}$.

$$\begin{array}{cccccccccc} \mathtt{b} & \not\succ & & & & \succ & & & \mathtt{d} & \not\succ \\ & & \not\succ & \mathtt{f} & \mathtt{g} & \succ & \mathtt{a} & \not\succ & & \\ \hline \mathtt{b} & & & \mathtt{f} & \mathtt{g} & \succ & \mathtt{a} & & \mathtt{d} & \end{array}$$

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

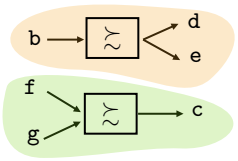


Figure 4.3: Covering scheme: a visual representation of Ex. 4.12

*"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 $\text{♀}_x \in N$ 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 $N := \{\text{♀}_x^1, \text{♀}_x^2, \text{♀}_x^3, \text{♀}_x^4, \text{♀}_x^5\}$ with the following preferences

$$
\begin{array}{ll}
\text{♀}_x^1\colon & a \succ_1 b \succ_1 f \succ_1 e \succ_1 c \succ_1 d \\
\text{♀}_x^2\colon & e \succ_2 b \succ_2 c \succ_2 d \succ_2 a \succ_2 f \\
\text{♀}_x^3\colon & f \succ_3 a \succ_3 b \succ_3 d \succ_3 e \succ_3 c \\
\text{♀}_x^4\colon & d \succ_4 a \succ_4 c \succ_4 e \succ_4 f \succ_4 b \\
\text{♀}_x^5\colon & c \succ_5 e \succ_5 b \succ_5 f \succ_5 d \succ_5 a
\end{array}
$$

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 $\text{♀}_x \in N$ is allowed to express only a binary judgment on each candidate $x \in \mathbb{X}$, which is either 'approved according to ♀_x' or not. The approved subset of candidates $\mathcal{A}_{\text{♀}_x} \subseteq \mathbb{X}$ should be an upset for the preference relation $\succsim_{\text{♀}_x}$. Hence, there is no pair of candidates $x, x' \in \mathbb{X}$ where x is preferred to x' w.r.t. $\succsim_{\text{♀}_x}$, x' is approved by ♀_x 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: $\mathcal{A}_{\text{🙎}^1} := \{a, b, f\}$, $\mathcal{A}_{\text{🙎}^2} := \{e, b, c\}$, $\mathcal{A}_{\text{🙎}^3} := \{f, a, b\}$, $\mathcal{A}_{\text{🙎}^4} := \{d, a, c\}$, $\mathcal{A}_{\text{🙎}^5} := \{c, e, b\}$, 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 $\mathcal{A}'_{\text{🙎}^1} := \{a, b, f\}, \mathcal{A}'_{\text{🙎}^2} := \{e, b, c, d, a\}, \mathcal{A}'_{\text{🙎}^3} := \{\}, \mathcal{A}'_{\text{🙎}^4} := \{d, a, c\}, \mathcal{A}'_{\text{🙎}^5} := \{c\}$ and sufficient coalitions $\mathcal{S}'$ containing the coalitions $\{\text{🙎}^1, \text{🙎}^2\}, \{\text{🙎}^5\}$ 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^C_α or Φ^P_α 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 $\mathcal{A}_{\text{🧍}^1}, \ldots, \mathcal{A}_{\text{🧍}^5}$, 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
GOOD	a	👤^1	👤^1	👤^4
	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: $\langle(\text{👤}^1, b, \{a, b\}, e, \{d, e\}), (\text{👤}^2, c, \{b, c\}, d, \{d, f\}), (\text{👤}^4, c, \{a, c\}, e, \{e, f\})\rangle$

- 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 $\alpha^* : \mathbb{X}^* \to \{ \textsc{Good} , \textsc{Bad} \}$, 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:

$$
\begin{aligned}
&a \succ_1 f \succ_1 b \succ_1 e \succ_1 c \succ_1 y \succ_1 d \succ_1 x\\
&e \succ_2 b \succ_2 y \succ_2 c \succ_2 d \succ_2 a \succ_2 f \succ_2 x\\
&f \succ_3 a \succ_3 d \succ_3 b \succ_3 y \succ_3 x \succ_3 e \succ_3 c\\
&d \succ_4 a \succ_4 c \succ_4 e \succ_4 x \succ_4 y \succ_4 f \succ_4 b\\
&c \succ_5 y \succ_5 e \succ_5 b \succ_5 f \succ_5 x \succ_5 d \succ_5 a
\end{aligned}
$$

It is not possible to represent the assignment (x, $\textsc{Good}$) together with the reference assignment α. Thus, x is necessarily assigned to $\textsc{Bad}$ 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).

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.

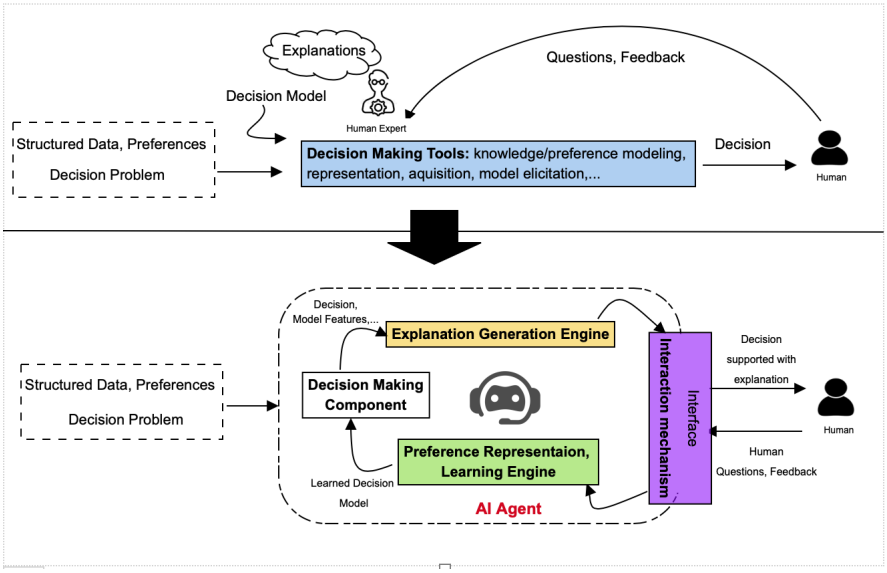


Figure 5.1: Dialectical vision for MCDA

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.

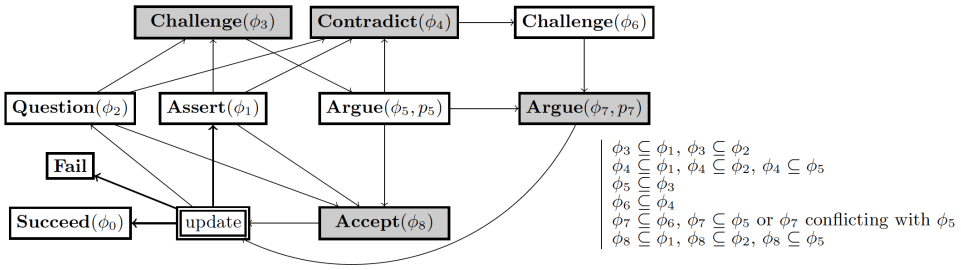


Figure 5.2: Successive speech acts at each iteration

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}.

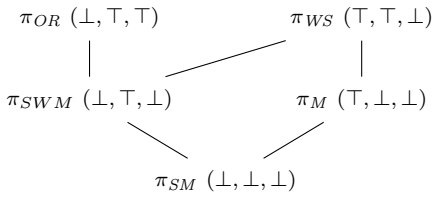


Figure 5.3: Structure $\mathcal{Q}$ with three properties

The 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.

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 $[c_1 = \textit{strong}]$, $[c_2 = \textit{strong}]$. 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 Example5.1, 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 Figure5.1) 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

Appendix A

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 $A = \{a, b, c, ...\}$ 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 $C^1, C^2, ...C^p$, 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, CentraleSupélec - 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 26^th 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.

Appendix B

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 28^{th} 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 27^{th} 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 26^{th} 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 30^{th} 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 19^{th} 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 19^{th} 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 14^{th} 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 14^{th} 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 20^{th} 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

2^{nd} 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 5^{th} 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 19^{th} 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 2^{nd} 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 9^{th} 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 1^{st} 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 4^{th} 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 4^{th} 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 25^{th} 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 22^{nd} 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 22^{nd} 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 77^{th} 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 77^{th} 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 20^{me} 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 20^{me} 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