Towards accountable decision aiding : explanations for the aggregation of preferences

Khaled Belahcene

université PARIS-SACLAY

CentraleSupélec

NNT : 2018SACLC101

Thèse de doctorat

Towards accountable decision aiding: Explanations for the aggregation of preferences

Thèse de doctorat de l'Université Paris-Saclay
préparée à CentraleSupélec

Ecole doctorale n°573 Interfaces - Approches interdisciplinaires: Fondements, Applications et Innovations (INTERFACES)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Gif-sur-Yvette, le 5 décembre 2018, par

KHALED BELAHCENE

Composition du Jury :

Nicolas Sabouret Professeur des Universités, Université Paris-Saclay (LRI)	Président
Denis Bouyssou Directeur de recherches, Université Paris-Dauphine (LAMSADE)	Rapporteur
Henri Prade Directeur de recherches, Univeristé Paul Sabatier (IRIT)	Rapporteur
Nic Wilson Senior Research Fellow, University College Cork (Insight)	Rapporteur
Sébastien Destercke Chargé de recherches, Université de Technologie de Compiègne (HeudiaSyc)	Examinateur
Christophe Labreuche R&D Engineer, Thales Research & Technology	Invité
Nicolas Maudet Professeur des Universités, Sorbonne Université (LIP6)	Co-directeur de thèse
Vincent Mousseau Professeur des Universités, Université Paris-Saclay (LGI)	Directeur de thèse
Wassila Ouerdane Maîtresse de conférences, Université Paris-Saclay (LGI)	Co-encadrante de thèse

Résumé

Nous cherchons à équiper un processus d'aide à la décision d'outils permettant de répondre aux exigences de redevabilité. Nous nous plaçons dans un cadre dialectique, mettant en scène deux protagonistes. Le *décideur* est amené à arbitrer entre des points de vue conflictuels. Un *analyste*, chargé d'éclairer la prise de décision, interroge le décideur quant à ses préférences—les compromis qu'il juge acceptable—jusqu'à obtenir une vision suffisamment claire de la situation pour pouvoir émettre une recommandation. Cette situation est souvent modélisée en termes d'agrégation de préférences reflétant des points de vue multiples, et résolue en faisant appel à un raisonement inductif, permettant à l'analyste de deviner au mieux les compromis entre ces points de vue qui satisfont le décideur. Une abondante littérature est consacrée à l'élicitation des préférences, où le paradigme dominant consiste à faire l'hypothèse d'un *modèle* paramétrique destiné à refléter l'attitude du décideur, puis à calibrer les valeurs des paramètres sur la base d'informations fournies par le décideur, à l'aide d'outils algorithmiques provenant de la Recherche Opérationnelle ou de l'Apprentissage Automatique. Afin de pouvoir rendre des comptes à des tierces parties, il nous semble pertinent de modifier ce paradigme, en prenant structurellement en compte le fait que l'information fournie par le décideur au cours du dialogue est *incomplète*, et en nous orientant vers l'élicitation dite *robuste*, où les recommandations sont inférées, non pas sur la base d'un unique jeu de paramètres jugé représentatif, mais sur l'ensemble des jeux de paramètres compatible avec l'information sur les préférences. Ainsi, les recommandations robustes, dans le cadre d'un modèle d'agrégation donné, sont déduites des éléments dialectiques. Afin de mettre en évidence ce lien déductif, sous la forme d'une *explication*—si possible correcte, complète, facile à calculer et à comprendre—nous nous sommes intéressés à la résolution d'un *problème inverse* concernant le modèle.

Dans le cadre de cet ouvrage, nous avons considéré deux formes de représentation du raisonnement: l'une ayant trait à la comparaison de paires d'alternatives fondée sur un modèle de valeur additive, l'autre ayant trait au tri des alternatives dans des catégories ordonnées fondé sur un raisonnement non-compensatoire.

La première partie de cet ouvrage est consacrée au modèle additif, qui joue un rôle central en décision multicritère, en choix social ou en apprentissage automatique. Nous y décrivons la procédure d'agrégation multicritère inférée dans le cadre de l'élicitation robuste dans le paradigme de la *représentation du raisonnement et des connaissances*, où l'information sur les préférence est représentée sous la forme d'une base de connaissance indépendante des requêtes, et où les requêtes sont évaluées par un moteur d'inférence. Nous décrivons dans ce paradigme un certain nombre de variantes du modèle additif, soit davantage contraintes, soit permettant un langage plus expressif pour les préférences. Nous présentons un cadre permettant de calculer la totalité de la relation de préférence nécessaire. Nous proposons deux moteurs d'explication pour le modèle additif. L'un est fondé sur un principe d'annulation des arguments, et permet de relier les prémisses, issues du dialogue, à une conclusion permettant de trancher en faveur d'une alternative. Nous en proposons diverses représentations, en établissons la correction et la complétude, et en discutons la simplicité, et montrons que le calcul d'explications simples est NP-complet. L'autre est fondé sur le principe de *diviser pour régner*, et consiste à décomposer une préférence complexe en une série de préférences simples enchainées par transitivité. Ce schéma est correct, mais nécessite de trouver un compromis entre complétude et simplicité. En outre, nous démontrons qu'il possède d'excellentes qualités en termes de complétude, de simplicité et de calculabilité lorsque l'on fait une hypothèse classique portant sur l'information sur les préférences.

La seconde partie de cet ouvrage est consacrée au tri non-compensatoire : il s'agit de porter un jugement sur la qualité intrinsèque des alternatives, en employant un langage restreint, ce qui limite la possibilité de trouver un compromis entre les différents points de vue. Nous commençons par proposer un cadre unifié permettant de décrire diverses variantes courantes de cette approche. Nous étudions le problème inverse consistant, étant donné un ensemble d'exemples d'affectations, à trouver une valeur du jeu de paramètre du modèle de tri non-compensatoire permettant de reproduire cette instance, et en proposons deux formulations en termes de satisfiabilité Booléenne, dont nous établissons la correction et la complétude. La première se fonde sur une description explicite de l'espace des paramètres, et requiert donc un nombre de variable exponentiel en le nombre de points de vue. Cependant, nous démontrons de

manière expérimentale qu'une telle formulation se montre environ cinquante fois plus rapide que les techniques fondées sur la programmation linéaire en nombres entiers qui constituaient l'état de l'art en la matière d'élicitation de modèles non-compensatoires fondés sur une règle majoritaire. A l'aide d'une caractérisation innovante des instances positives du problème inverse, nous contournons l'obstacle de la représentation explicite de l'ensemble des coalitions de points de vue, et proposons une seconde formulation, celle-ci compacte, avec un nombre polynomial de variables et de clauses. Nous discutons l'apport de ces techniques à l'élicitation robuste de modèles de tri non-compensatoire, dans une optique de parvenir à proposer un dispositif d'aide à la décision à même de rendre des comptes. Enfin, nous envisageons la contribution de ces techniques à l'élaboration d'une procédure de tri où des candidats seraient soumis à l'approbation d'un jury, et pour laquelle l'exigence de rendre compte des décisions est prise en considération de manière structurelle, dès la conception. Nous proposons des certificats corrects et complets permettant d'attester de la régularité procédurale, sous la forme de schéma d'arguments, bien que la simplicité ne soit pas garantie dans le cas général. Nous proposons aussi des certificats corrects, sous la formes de schémas d'arguments, attestant de la nécessité de certaines décisions, permettant de répondre à de potentielles contestations de décisions particulières.

Remerciements

Cet ouvrage est le fruit d'une intense et plaisante collaboration avec Christophe Labreuche, Nicolas Maudet, Vincent Mousseau et Wassila Ouerdane. Ces quatre figures tutélaires ont nourri ma réflexion et mon estomac pendant plus de trois ans et, ensemble, nous avons formé une armée mexicaine aussi joyeuse que redoutable. *Caramba !*

Je remercie aussi toutes les personnes avec qui j'ai partagé cette séquence de travail : de manière régulière, au laboratoire Génie Industriel de CentraleSupélec, et en particulier sa "cinquième équipe", ou au laboratoire d'Informatique de Paris 6, ou de manière plus épisodique, les membres des communautés de l'Aide à la décision multicritère et de l'Intelligence Artificielle fondamentale, que je retrouve avec plaisir au gré des séminaires et colloques. En particulier, j'ai eu la chance de collaborer de manière fructueuse avec Marc Pirlot, Olivier Sobrie, Yann Chevaleyre, Antoine Rolland, Patrick Meyer, Alexandru Olteanu, Oumeima Khaled, Ali Tlili, Sébastien Destercke, Jean-Marie Lagniez et Fabien Tarissan. Mes remerciements vont aussi à Ulle Endriss, Patrice Perny, Abdallah Saffidine et Anaëlle Wilczynski.

Enfin, je remercie tout particulièrement les rapporteurs de cet ouvrage, Denis Bouyssou, Henri Prade et Nic Wilson, qui ont accepté de lire cet ouvrage de manière approfondie, et qui ont contribué à sa qualité par leurs remarques pertinentes et constructives.

Par ailleurs, ce projet n'aurait jamais pu voir le jour sans le soutien tendre, joyeux, quotidien et sans faille de mes colocataires. Cet ouvrage leur est dédié[a].

[a]En revanche, il doit être noté que cet ouvrage ne doit pas grand chose à la ville de Redon ni au RER B.

À Annaïck, Leïla, Sarah et Maël.

Contents

Exposition **1**

1 Positioning 3
1.1 Decision Aiding . . . 3
1.1.1 The decision aiding process . . . 3
1.1.2 Accountability . . . 5
1.2 The elicitation process . . . 8
1.2.1 High-level description . . . 8
1.2.2 The aggregation model . . . 10
1.2.3 Using preference information to specify a model . . . 13
1.3 What we propose . . . 17
1.3.1 Explaining the reasoning itself, not its product . . . 17
1.3.2 A dialectical take to preference information . . . 18
1.3.3 Robust elicitation as deductive reasoning . . . 21
1.3.4 Explaining robust adjudications . . . 22
1.3.5 The shape of things to come . . . 25
1.4 Formalization . . . 26
1.4.1 Points of view and alternatives . . . 26
1.4.2 Problem type . . . 28
1.4.3 Aggregation models, the inverse problem and necessary adjudications . . . 29

Part I : Comparing with an additive model **31**

Foreword 33
Comparing alternatives with additive values . . . 33
Usage of the additive value model . . . 34
Research questions . . . 36
Chapters . . . 37
Featured contributions . . . 38

2 Robust elicitation as reasoning 39
2.1 Introduction . . . 39
2.1.1 State of the art . . . 39
2.1.2 A grim situation . . . 41
2.1.3 A motivating example . . . 42
2.1.4 To the rescue: Farkas' lemma . . . 43
2.1.5 Towards an interpretable framework for necessary preference . . . 44
2.2 A toolbox for the inference of necessary preferences . . . 45
2.2.1 Core values and alternatives . . . 45
2.2.2 Unbounded pairs . . . 46
2.2.3 Core intervals and indexes . . . 46
2.2.4 Orientation of a core interval w.r.t. a pair of alternatives 46
2.2.5 Covectors operating on differences of value . . . 47
2.2.6 Representation of dominance through covectors . . . 47
2.2.7 Representation of preference through covectors . . . 47
2.2.8 Inference of necessary preference through covectors . . . 49
2.3 A working example . . . 49
2.3.1 Inputs . . . 50
2.3.2 The encoder . . . 52
2.3.3 The inference engine . . . 55
2.3.4 Outputs . . . 56
2.4 What we have done and why it is important . . . 58
2.4.1 A flexible framework . . . 58
2.4.2 A streamlined representation . . . 60
2.5 Perspectives and venues for improvement . . . 63
2.5.1 Relaxing orientations to account for interpolation . . . 63
2.5.2 Encoding the intensity of preferences . . . 68

3 Explanations 71
3.1 Explaining via cancellation . . . 72
3.1.1 Cancellative properties of the additive value model . . . 72
3.1.2 Syntactic cancellation . . . 73
3.1.3 Elliptic cancellation . . . 79
3.1.4 Computing cancellative explanations . . . 84
3.2 Explaining via a sequence of preference swaps . . . 90
3.2.1 Sequences of *even-swaps* . . . 91
3.2.2 Explaining with a sequence of *preference-swaps* . . . 94
3.2.3 Properties of the *preference-swaps* explanation engine . 95
3.2.4 Results in the case of a binary core . . . 97

Part II : Sorting with a Noncompensatory model **105**

Foreword 107
Sorting into ordered categories . . . 107
Noncompensatory models of preferences . . . 107
Applications of Noncompensatory sorting . . . 108
Research questions . . . 109
Chapters . . . 109
Featured contributions . . . 110
See also . . . 110
Noncompensatory sorting without frontiers . . . 111
Comparing the SAT formulations for Inv-NCS . . . 111
A SAT formulation for noncompensatory ranking with RMP . 111

4 The noncompensatory sorting model 113
4.1 The noncompensatory sorting model (NCS) . . . 113
4.1.1 Sorting with a single profile. . . . 113
4.1.2 Sorting into multiple categories . . . 114
4.1.3 Interpretation of the aggregation procedure . . . 114
4.2 A working example . . . 115
4.3 Variants . . . 118
4.3.1 Presentation with profiles . . . 118
4.3.2 Limited array of parameters. . . . 119
4.3.3 k-additive representations of sufficient coalitions . . . 119
4.3.4 Description in other paradigms. . . . 120
4.3.5 Veto. . . . 121

5 SAT-based formulations for Inv-NCS 123
5.1 Introduction . . . 123
5.1.1 NCS and Inv-NCS . . . 123
5.1.2 Encoder and decoder. . . . 125
5.1.3 A tale of two formulations. . . . 125
5.2 A SAT formulation for Inv-NCS based on coalitions . . . 125
5.2.1 Informal presentation of the approach . . . 126
5.2.2 A SAT encoding of an instance of Inv-NCS . . . 127
5.2.3 From a solution of Inv-NCS to a solution of the SAT formulation . . . 128
5.2.4 Decoding a solution of the SAT formulation into NCS parameters . . . 129
5.3 A SAT formulation based on pairwise separation conditions . . 130
5.3.1 Inv-NCS with fixed approved sets . . . 131

5.3.2 A pairwise characterization of positive instances 133
5.3.3 A compact SAT formulation for Inv-NCS 133
5.3.4 More than two categories 135
5.4 Computational complexity of Inv-NCS 137
5.4.1 Complexity of Inv-NCS with fixed approved sets 137
5.4.2 NP-hardness of Inv-NCS 138

6 Experimental results concerning for Inv-NCS via SAT 141
6.1 Introduction 141
6.1.1 Context 141
6.1.2 Research question 142
6.1.3 Layout of the chapter 143
6.2 Learning MR-Sort using Mixed Integer Programming 143
6.3 Implementation 146
6.3.1 Experimental protocol and implementation details . . . 147
6.3.2 Intrinsic performance of the SAT-C formulation 149
6.3.3 Comparison between the formulations 150
6.4 Discussion and perspectives 155
6.4.1 Influence of the parameters 155

7 Accountable decisions with Inv-NCS 161
7.1 Robust elicitation of a noncompensatory sorting model 161
7.1.1 The situation 161
7.1.2 State-of-the-art approaches 162
7.1.3 Contributions of Inv-NCS 165
7.2 Accountable Approval Sorting 169
7.2.1 The context: selection by a jury 169
7.2.2 Accountability requirements 171
7.2.3 Addressing overall *Procedural regularity* with Inv-NCS . 173
7.2.4 Contestability of individual decisions 176

Conclusion **185**

8 Conclusion 187
8.1 Summary of our contributions 187
8.2 Open questions and work in progress 189
8.3 Perspectives 190

Bibliography 195

Appendices **211**

A Proofs 213
A.1 Proofs of Theorems 2.4 and 2.8 213
A.2 Proof of Theorem 3.9 218
A.3 Proof of Theorem 3.10 220

Supplementary material **225**

B Accountable sorting without frontiers 227

C Comparing the SAT formulations for Inv-NCS 237

D Learning parameters of RMP with a SAT formulation 247

Exposition

1

Positioning

1.1 Decision Aiding

We consider a *decision aiding process*, as described in [Bouyssou et al., 2006, Tsoukiàs, 2008], occurring between somebody looking for decision support, and an analyst providing such a support. We assume the decision maker and the analyst have decided to follow a *principled* approach to decision aiding, based on an *evaluation model* [Bouyssou et al., 2000].

1.1.1 The decision aiding process

In order to settle on a course of action, the decision maker and the analyst engage in a dialectical process. Questions and answers are exchanged, and at the end, the decision maker should emerge with a vision of the situation clear enough to permit an enlightened decision making.

Final outcome. The stop condition of this dialog is not perfectly clear. As forcefully formulated as the concept of bounded rationality by Simon in his analysis of decision processes in organizations [Simon, 1991], it cannot be expected that the decision eventually reached at the end of the process is perfectly justified, rational or optimal in a strong sense. Decision processes, and, a fortiori, decision aiding processes, use limited resources—time and cognitive capabilities of the decision maker and the analyst. Therefore, what can be expected is that the recommendations formulated at the end of the process, which are issued from necessarily imperfectly determined models, should be

considered as convincing by the decision maker. In particular, they[a] should be convinced that the solutions emerging from the process are satisfactory. Following [Rawls, 1971], we call *considered judgement* the final, stable and desirable point of the decision aiding process, and we note that [Cailloux and Meinard, 2018] is a recent attempt at transposing this notion in the dialectical framework of a decision aiding process, as a *deliberated judgment*.

Elicitation of an evaluation model. So as to converge towards this promised end, the protagonists often decide to follow a *principled* approach to decision aiding, based on an *evaluation model* [Bouyssou et al., 2000]. An evaluation model is a mathematical construct that allows to describe the situation in an unambiguous manner, and should yield a recommendation about the course of action the decision maker has to settle upon. Besides these *descriptive* and *prescriptive* functions, we believe the evaluation model plays an important *constructive* role. By abstracting away from the situation, and maybe moving slightly away from the actual options of the decision maker, it establishes a fiction that may help the decision maker to make their mind about the adjudication of trade-offs between the various points of view. The analyst can take advantage of this induced detachment to carefully *elicit* some insights about the preferences of the decision maker, then incorporated them into the evaluation model that eventually yields the sought answer.

The science of decision aiding This type of decision aiding process has been an object of much scrutiny for the last fifty years. At the crossroads between Operations Research, Economy, Applied Mathematics, Psychology, Management Science and Computer Science, *Decision Theory* is interested in providing formal tools to describe, analyze, equip and enhance decision processes and decision aiding processes. Particularly, decision situations involving:

a) several points of view ; and/or

b) several decision makers ; and/or

c) partial or imprecise information about the consequences,

are known to require specific efforts to model and assist. While there is only one decision maker in the situation we address, we assume that multiple points of view should be taken into account; therefore, the problem falls under the umbrella of *multiple criteria decision aiding*. The evaluation model needs to adjudicate trade-offs between conflicting points of view; as there is no canonical

[a] We use *singular they* as a gender-neutral pronoun. This allows us to eschew the awkward 'he or she', or the even more awkward assignment of each protagonist to a gender.

way of doing so, it needs to account for the *preferences* of the decision maker, by obtaining some kind of *preference information*.

1.1.2 Accountability

We consider the decision aiding process being subject to multiple demands for accountability.

The decision maker. Accountability is first and foremost due to the decision maker. They were looking for decision support, and expect the analyst to be sincere and trustworthy, and help them in reaching a considered judgment. This demand is even more urgent, should we consider the role of the analyst is played by an artificial agent—which is exactly what *recommender systems* [Ricci et al., 2010] try to do—as the designer of such systems shall endorse full responsibility if the recommendation is biased or insincere. We translate this generic demand for accountability into a (non exhaustive) list of requirements for the decision aiding process.

- *Adequacy:* obviously, the analyst needs to design an adequate evaluation model, and to elicit it adequately. While there are many scientific articles that may offer a solid basis to address the second point, the first point is really difficult, as illustrated e.g. by [Condorcet, 1785, Bouyssou et al., 2000].
- *Sincerity w.r.t. limited knowledge:* many elicitation methods proposed in the scientific literature either disregard mundane constraints such as time, patience or precision, or propose inference techniques that go beyond the information provided by the decision maker to calibrate their model. When preference information is incomplete, the analyst should take great care in determining whether their recommendations hold whatever the completion might be, or if they are making an educated guess in order to complete it, that might be falsified if additional information would become available. This epistemic state should be reflected in their attitude when presenting the recommendation to the decision maker. The methods used in the decision aiding process should therefore integrate the unavoidable imperfect character of the information (evaluations, preferences, etc.) and of the models. The analyst should make every effort to produce robust recommendations i.e. recommendations that are prudent, flexible, stable, in view of the imperfections of the information and the models.
- *Behavioral effects:* behavioral sciences (broadly conceived) show how sensitive people are to apparently irrelevant or non-significant details of their environments when making a choice: how information is displayed,

which words are used to frame the options and their consequences, etc. The analyst should be aware of most of these effects, and do their best to protect the decision maker from these, as they might constitute alienating obstacles preventing to attain a considered judgment.

- *Empowerment:* the decision aiding dialog is rarely unstructured. The analyst often leads the process, and questions the decision maker according to an agenda where the goal is to rapidly elicit the model. Practical applications of this decision aiding framework sometimes report the need for going back and modify elements previously constructed—adding or removing points of view, modifying the objects of the evaluation model, or the type of results yielded by the evaluation model—but backtracking is rarely considered in the scientific literature concerned with evaluation models. Steps could profitably be taken towards a more mixed initiative. In particular, instead of being assigned to a purely passive role, the decision maker could be allowed two important dialog acts: 'why?', requesting an *explanation* supporting a claim made by the analyst, and 'no!', expressing an explicit disagreement with such a claim, and prompting a backtrack of the elicitation process.

Stakeholders of the decision. The decision aiding process may also consider the need to account for the recommendation made to stakeholders of the decision that have not been involved in the decision. The answer to this issue can be thought of as the provision of an *explanation*, that conveys additional information complementing the recommendation. These explanations may aim at addressing two separate issues:

- maybe the stakeholders are allowed to contest the recommendation, and the request of an explanation is actually a challenge to the validity of the recommendation;

- maybe the stakeholders require to better understand the basis of the recommendation, in order to better anticipate the behavior of the decision making entity during future interactions.

Society at large. Similarly to Operations Research, Decision Aiding has been chiefly aimed at helping chief executives to take 'better' decisions—more efficient in the case of OR, better reflecting their priorities in the case of DA. This aim was well aligned with the role of the analyst, often played by a consultancy firm, reserving decision aiding for high-impact decisions.

Now that computers have become ubiquitous, it seems desirable to give

access to decision aiding to the proverbial [b] 'person in the street', as we are all decision makers, when it comes to opting for a career, a holiday destination or a new coffee machine. A challenge in designing an artificial agent capable of decision aiding consists in obtaining the same level of trust that a human analyst, sanctified by formal and practical training in decision aiding matters, but also capable of sympathizing to the concerns of his fellow human decision maker, may benefit from. *Accountability* of the decision process is thus a key enabler on the path towards its potential automation. This demand is made all the more urgent when considering the rising defiance for 'Artificial Intelligence', that conflates the dreads of ubiquitous surveillance, manipulation, depowerment and unemployment.

Even in the case where decision aiding remains human based, thus reserved for high-stakes situations, the willingness of the public to submit themselves to decision from above—were they taken by *enlightened despots*, in the sense of Voltaire—might have receded since the inception of OR in a world polarized by global conflicts. It seems that people, rather than blindly following their so-called 'elites', demand more accountability from them. Any person in the position of making a decision should be prepared to face the question 'why should I trust you?' Note that the purpose of accountability is to answer distrust, which is definitely not the same as promoting trust[c]. Distrust, or defiance, is an exogenous state of affairs, that we need to cope with[d].

This trend towards a more 'horizontal' society, and the rising concerns about automated decision aiding, are somehow reflected in the legal and regulatory systems of western societies. For instance, the *General Data Regulation Policy* (see e.g. [Goodman and Flaxman, 2017, Wachter et al., 2017]), at the European Union level, or the *Loi pour une République Numérique* in France (see e.g. [Besse et al., 2017]), are recent regulatory texts that try do define, and enforce accountability, even though both the novelty of the so-called 'right for an explanation', and its legally binding aspect, are questionable: what somehow defines a society under the rule of law is both the imperious necessity, for a government, to duly motivate its decisions, and the right given to individuals to contest them, even though the actual accountability of the government towards the governed very much depends on the type and content of the explanation required to support a decision.

[b] but gender-free

[c] [Tintarev, 2007] identified the promotion of trust as one of the possible goals of an explanation, in the context of recommender systems (see e.g. [Rossi et al., 2011]), where the quality of the *user experience* is a key factor for (commercial) success.

[d] It might be considered as negative—a nihilistic force hampering every action, or positive—a manifestation of the people's free will

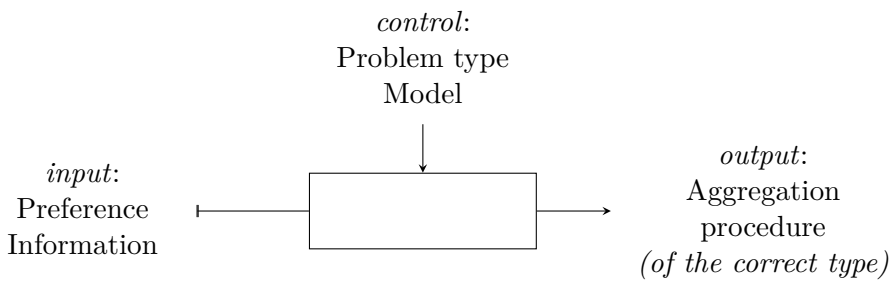

Figure 1.1: The elicitation process.

1.2 The elicitation process

At some point during the decision aiding process, a decision maker and an analyst following a principled approach should engage in an *elicitation process*, aiming at building an evaluation model that should reflect the view of the decision maker and help them in the resolution of their dilemma.

This process can be implemented in many ways. In this section, we give a high-level description of the process, and quickly review some common approaches, with a particular attention to accountability issues.

1.2.1 High-level description

We give a high-level description of the elicitation process: the context it needs to operate, the inputs it operates upon, and the outputs it yields. Figure 1.1 illustrates this description. We detail the information that should be provided at the onset of the process, the information that shall be processed during elicitation, and the expected output of the process, together with their requirements.

Aggregation procedures. The elicitation process is expected to yield an *aggregation procedure*, whose role is to bring together and combine a multiplicity of points of view into a single overall judgment. The aim of the elicitation process is to yield an aggregation procedure that: i) reflects the views of the decision maker, and ii) helps them solve their dilemma. The second point is tied to the *type of problem* the aggregation procedure is designed to solve, e.g.:

- *sorting* problems consists in assigning alternatives to categories, known in advance and ordered by level of requirement;
- *pairwise comparison* problems consists in adjudicating, between two alternatives, which one is the fitter;

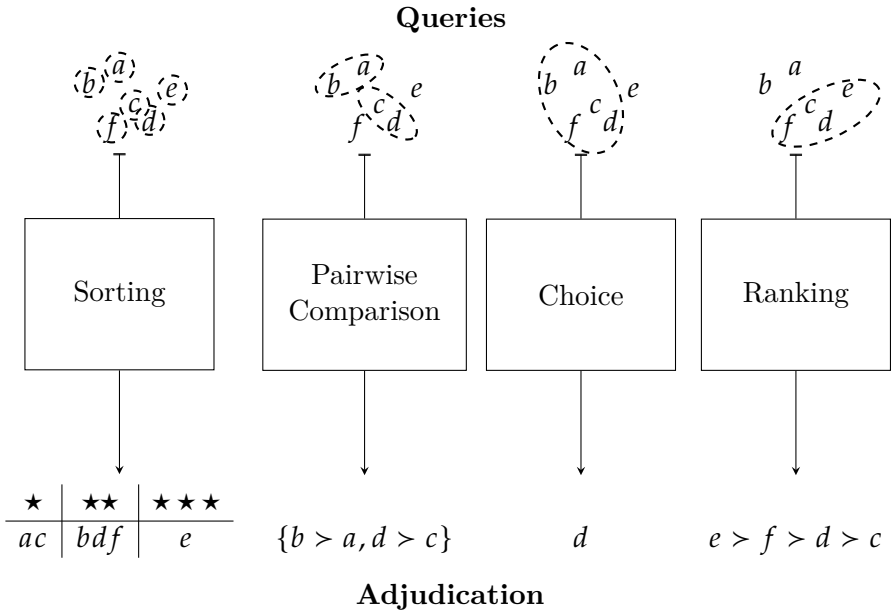

Figure 1.2: Aggregation procedures.

- *choice* problems consists in selecting the fittest alternative among any group;
- *ranking* problems consists in ordering any group of alternatives from least fit to fittest.

Given a problem type, we call *queries* the potential inputs of an aggregation procedure, and *adjudications* its potential outputs. Aggregation procedure of various types are depicted on Figure 1.2.

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 provided to the elicitation process. They are usually defined in a prior phase, sometimes called *problem structuring* [Bouyssou et al., 2000].

Obtaining, representing and mechanizing the views of the decision maker. Trying to faithfully represent the views of the decision maker is a challenging endeavor, that should not be underestimated. The problem of *aggregating preferences* has been the chief concern of the Social Choice domain for more than two centuries: how should the individual preferences of a population of agents be aggregated into a *social* preference? When the situation is consensual—i.e. when an alternative is preferred to another according to all points of view, the social preference is clear. Unfortunately, it is often necessary to arbitrate between conflicting points of view. Therefore, the function of an

aggregation procedure is precisely to adjudicate trade-offs between points of view. For a long period of time, *normative* approaches to Social Choice have tried to uncover or defend supposedly 'good' aggregation procedures, but the discussion between Condorcet and Borda in the 1780s had already unveiled paradoxes in seemingly 'reasonable' aggregation procedures [Condorcet, 1785], and the generality of these conundrums has been established in the 1950s by Arrow's theorem [Arrow, 1950]: there are a number of desirable properties for the aggregation procedure, that cannot be satisfied simultaneously. Consequently, there is none 'universally good' aggregation procedure, only a large set of imperfect ones that are more or less adequate to a given situation. The stakes of the elicitation process reside in sculpting an adequate aggregation procedure, with a reasonable amount of efforts.

Preference information. *Preference information* encompasses any information provided by the decision maker to the elicitation process. It is the raw material processed during the elicitation of the aggregation procedure. The questions concerning preference information organize the elicitation process:

i) What type of preference information should be obtained?

ii) How to collect preference information?

iii) How preference information should be processed so as to sculpt the aggregation procedure?

iv) How to account for imperfect preference information?

All these questions need to be considered carefully, and there are many sensible options available to address each one of them.

The type of preference information, and its usage in the specification of the aggregation, shall be discussed after we present a device playing a key role in the elicitation process: the *aggregation model*.

1.2.2 The aggregation model

The *principled* approach to decision aiding advocates the use of an *aggregation model* to funnel the elicitation process. While the aim of the elicitation process is to obtain, represent and mechanize the views of the decision maker, the aggregation model supports this aim in many ways.

Structuring the elicitation process. Technically, an aggregation model consists in a parameterized family of aggregation procedures, which can be considered as a partially specified aggregation procedure. Each value of the *preference parameter* specifies a single aggregation procedure. Therefore, the

goal of the elicitation process is to interpret the preference information so as to pinpoint the value of the preference parameter, so as to yield the corresponding procedure. Models can be considered as frames providing structure for the elicitation process, channeling the effort of the decision maker and the analyst into incrementally sculpting the aggregation procedure.

Languages for preference. Following [Perny, 2000, Rolland, 2013, Grabisch and Labreuche, 2010], the aggregation models can be sorted in three families, according to the language used to describe the aggregation procedures, to which we append a fourth family encompassing all the models that deliberately circumvent an explicit representation.

- *Aggregate, then compare*: the procedures aim at computing an overall numeric score, the *value*, representing the overall fitness of an alternative, then the usual ordering of numbers is used to sort, compare, choose or rank alternatives.
- *Compare, then aggregate*: the procedure is analogous to a society where preferences according to each points of views need to be aggregated into an *outranking relation* denoting overall preference, then this relation is *exploited* to yield an answer permitting to compare, sort, choose or rank alternatives.
- *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, the ...' have been used to formally describe preferences for a long time, e.g. *expert systems* [Waterman, 1986] implementing decision trees, or the *dominance-based rough set approach* [Greco et al., 2001b]; more sophisticated languages for representing preferences have been proposed, e.g. based on logics [Rossi et al., 2011, Kaci, 2011], or on representations of conditional dependencies between points of view based on graphs[e], such as *generalized independence networks* [Gonzales and Perny, 2004], *conditional preference networks* [Boutilier et al., 2004], or *conditional possibilistic preference networks* [Amor et al., 2015]—see e.g. [Amor et al., 2016] for a recent survey.
- *transductive approaches*: the recommendations are obtained by establishing a correspondence between the particular case at hand (the query) and some particular facts of the knowledge base, without paying attention to the explicit representation of the aggregation function. While *case-based reasoning* has not yield significant developments in dealing with

[e]These representations are often referred to as *compact*, as they allow for a somewhat factorized representation of dependencies.

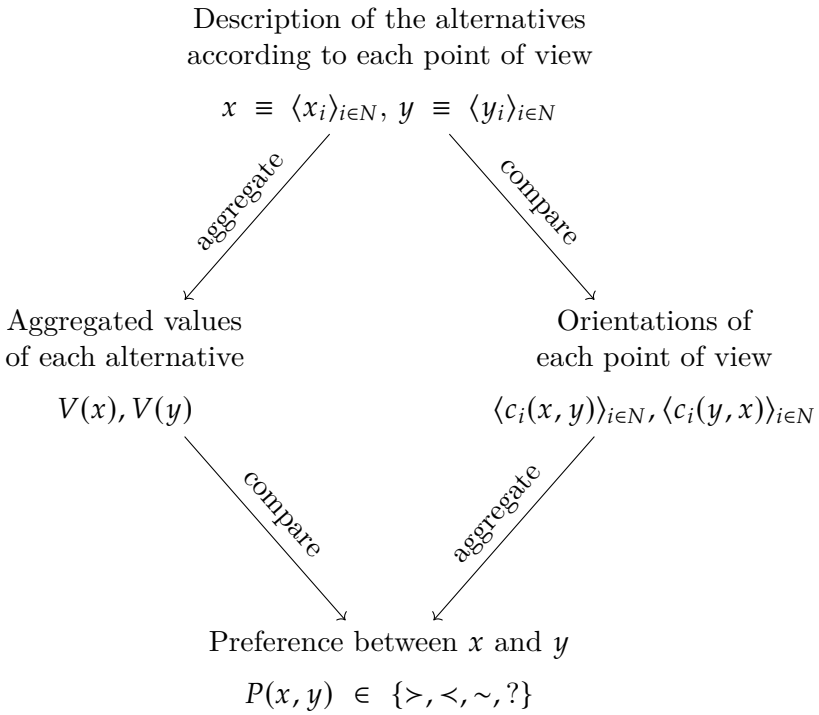

Figure 1.3: Diagram for the aggregation of preferences: compare then aggregate, or aggregate then compare?

preferences, there are several approaches for reasoning with preferences based on *analogy*—e.g. [Prade and Richard, 2018, Bounhas et al., 2018], and others based on *possibilistic logic*—e.g. [Dubois et al., 2005, Gérard et al., 2007]. While this report assumes an explicit representation of aggregation procedures, some explanatory devices proposed in Chapters 3 and 7 are naturally described in the transductive paradigm.

The value-based and outranking-based approaches to modeling preference are illustrated, in the case of a pairwise comparison problem, by Figure 1.3.

Choosing a family of models largely determines the language used during the elicitation dialog.

Theoretical properties. Models can be considered as normative stances towards decision making. Theoretical studies aimed at describing the properties of aggregation procedures are abundant in Social Choice [Brandt et al., 2016]. Conversely, studies aimed at, given a particular decisional stance, describing the corresponding *representation*—the set of the aggregation procedures that satisfy this stance—has been the focus of many studies in Economy and

Conjoint Measurement [Debreu, 1960, Fishburn, 1976, Krantz et al., 1971].

Selecting a model. The first step, in such a principled approach, is to select a model. This is an important decision conditioning the rest of the elicitation process. Guidelines and principles about this question can be found in [Roy and Słowiński, 2013, Bouyssou et al., 2000, Guitouni and Martel, 1998]. We believe this question has bearings on the accountability of the decision aiding process, as some models might yield an inadequate reflection of the views of the decision maker, but, throughout this report, we adopt a normative attitude towards the choice of the model. For us, the model is a provided element that should not be discussed—but we shall see that our results might fuel this discussion with arguments grounded in tangible elements. In a sense, our position is close to [Cailloux and Endriss, 2016], where, in a Social Choice context, in the position of deciding on a specific social rule (an *aggregation procedure*, in our parlance), discussions about axioms that should be satisfied by the social rule in the abstract should be avoided, but rather be instantiated through prototypical examples.

1.2.3 Using preference information to specify a model

Once a model is selected, efforts should be made to collect preference information to select a value for the preference parameter. Several approaches can be, and have been, advocated:

Direct elicitation. A straightforward approach to the elicitation problem would consist in promoting a discussion bearing directly on the preference parameter. We have to forcefully object to this option:

- when taking place between experts in Decision Theory, this normative discussion is often barren—this is a consequence of Arrow's theorem;
- a preference parameter has no intrinsic value, and should never be considered outside the scope of the model it parameters; for instance, several models refer to the 'importance of points of view', but models such as lexicographic orders, conditional preference networks, additive value, or weighted majority can interpret this statement in different manners. Communicating about the preference parameters seems like a sure way to create confusion and misunderstanding.

Consequently, while we endorse the use of an evaluation model, we prefer to consider it *latent*, and to limit the dialog to mentioning manifestations of the model, e.g. absolute or comparative evaluation of alternatives.

Preference information can be qualified as *holistic* when it is restricted to mention phenomena that are potentially observable—i.e. alternatives—rather than theoretical constructs of the model. The elicitation of an aggregation model based on such holistic preference information is dubbed *indirect*.

Complete indirect elicitation—Standard sequences. From a theoretical point of view, the first question to consider when indirectly eliciting a model concerns the *observability* of the parameters: is the parameter space one-to-one with the observations? If not, no indirect elicitation procedure is able to differentiate values of the preference parameter yielding similar results—see e.g. Example 7.1 in Section 7.1.2. This has, of course, no bearing on the outcome of the elicitation, but should raise alarms concerning the relevance of mentioning those parameters, as they are pure artifacts of the language chosen to describe the model.

A second, very important question concerns the provision of a constructive algorithm yielding a questioning strategy—a list of *holistic preference queries* so as to determine, up to an arbitrary precision, the value of the preference parameter, assuming the decision maker preferences are described by an aggregation procedure of the model.

Such a *standard sequence* represents a 'gold standard', and ensures the model is relevant for supporting a decision aiding process. Nevertheless, the practical interest of this platonic ideal is mitigated by some difficulties stemming from the human-in-the-loop nature of the querying process:

- the queries postulated by the standard sequences often involve fictitious alternatives with extreme attributes, that tend to stretch their plausibility to the point of absurdity;
- the decision maker has limited availability or patience. Therefore, the elicitation process should be designed in order to yield results even if it is interrupted sooner than expected;
- The information collected is imperfect. The decision maker may make blunders—either random or systemic, due to a cognitive bias—and not report their true preference. They may also change their mind during the elicitation process, maybe because of the reflexive and constructive nature of the elicitation process.

Incomplete indirect elicitation—Disaggregation. To face these issues, a popular solution consists in adopting a *learning* approach. Preference information is considered as external *data*, provided *as is*, rather than as a (living and breathing) database that can be queried at will. Thus, the elicitation

process has to do with an input that is limited in length and quality, but hopefully meaningful.

Depending whether the focus is on the product or the manner, this process is sometimes dubbed as 'learning', or 'disaggregation':

- *learning*, because the final product is a compiled form of knowledge—in the form of an aggregation procedure—that can be efficiently used to address decision situations [Geffner, 2018];
- *disaggregation* [Jacquet-Lagrèze and Siskos, 2001], because it transforms holistic preferences information that supposedly derives from the aggregation procedure into information about the parameters governing this aggregation procedure.

Approaches that call themselves 'disaggregation' often proceed as follows: in a first phase, preference statements about alternatives are disaggregated—translated into statements about parameters; then, either the set of parameters compatible with these statements is empty, reduced to a singleton, or larger. If it is empty, either the analyst decides to extend the aggregation model, or they try to find a value of the parameter that 'best reflects' the statements of the decision maker. If the set is reduced to a single parameter value, then the elicitation is complete[f]. Should the set of compatible parameters contain more than one element, either more preference information is collected, or a specific value of the preference parameter is singled out from the set of values compatible with the preference information. Many methods heave been advocated to implement a choice function yielding 'the most representative preference parameter', hence, the 'most representative aggregation procedure'[g], and we briefly describe some of them.

- *Soft constraints:* the statements about parameters are formalized as constraints, and processed in an optimization framework, with the satisfaction level of the constraints as an objective; this method was made popular in multiple criteria decision aiding by [Siskos et al., 2005], and permits to represent and solve the *underconstrained* (with too many compatible parameters) and the *overconstrained* (with no compatible parameters) in the same framework, almost seamlessly.
- *Central tendencies:* median, arithmetic or geometric mean [Salo and Hamalainen, 2001], moment minimization [Bous et al., 2010]—depending on the operations meaningfully supported by the parameters of the model— may yield a value at the 'center' of the set of compatible parameters.

[f] Congratulations!

[g] See e.g. [Kadzinski et al., 2012].

- *Metrics of the parameter space:* e.g. Stochastic Multicriteria Acceptability Analysis [Lahdelma et al., 1998, Tervonen and Figueira, 2008] considers (normatively) a probability measure on the set of compatible models and computes indexes corresponding to relative volumes. It extends early works by [Charnetski and Soland, 1978] and computes expected values thanks to Monte Carlo simulation techniques.

- *Principled error minimization:* in Preference Learning [Fürnkranz and Hüllermeier, 2010], the probability measure on the parameter space is tied to a noise model, assuming a latent *ground truth*, accounting for the observed preference information. Selecting a parameter can then be achieved according to a fully Bayesian approach, or simply refer to maximum likelihood.

- *Entropy maximization:* generalizing Laplace's principle of insufficient reason, this estimation approach prescribes to aim for the less specific parameter.

Principled elicitation processes following the disaggregation approach can be described by the workflow presented in Table 1.1.

Step	Action
1	Collect holistic preference information, under the form of a set of queries, together with their adjudication. This knowledge base functions as a learning set for the problem of fitting the preference parameter.
2	Transcribe each piece of preference information into a set of constraints bearing on the value of the preference parameter.
3	Derive a value for the preference parameter from the resolution of an optimization problem constrained by the conditions transcribing the preference information.
4	Adjudicate new queries, not belonging to the learning set, by following the sorting procedure corresponding to the chosen model and the optimal parameter.

Table 1.1: The *disaggregation framework* for learning to adjudicate from assignment examples. *Adjudication* can be e.g. the assignment of an alternative to a category (for sorting problems) or the specification of pairwise preference between two alternatives (for problems based on comparisons, e.g. choice or ranking).

1.3 What we propose

In an effort towards the accountability of the decision aiding process, we propose to provide *explanations* shedding light not on the product of the process—the evaluation model elicited during the process, and supposedly representing the views of the decision maker—but on the elicitation process itself.

1.3.1 Explaining the reasoning itself, not its product

Since the launch of the XAI—for eXplainable Artificial Intelligence— by the Defense Advanced Research Projects Agency in 2017 [Gunning, 2017], many studies have focused on the task of explaining a *learner*. In this work, we challenge the applicability of this approach to a principled decision aiding process along two angles:

i) According to the *constructive* approach to the decision aiding process advocated by [Roy, 1993], there is nothing like a definite preference object in the decision maker's mind preexisting the decision aiding process. This comes in contradiction with the (statistical) learning paradigm based on the notion of a latent ground truth. It also hints at a possible inadequacy of the *didactic* paradigm of explanation usually and implicitly assumed, with an asymmetric role assigned to the explainer and the explainee:
   - when submitted to a validation process and asked the question 'why should I trust you' [Ribeiro et al., 2016], the artificial agent is in the seat of the pupil trying to give the right answer to the teacher;
   - once validated and put in production, the recommender system asked by a user to explain its decision is expected to provide an answer that is often not aimed at reconsidering the recommendation, but at teaching the user to better live with it[h].

ii) A learner is structurally geared towards compiling vast amounts of knowledge into an efficient procedure. While this approach is relevant for repeatable tasks, both in terms of available data and of opportunity to use it again, we believe decision aiding processes do not fit this description. Each process is singular, hopefully rather short, and inherently more concerned with accountability than with efficiency.

Therefore, we advocate:

i) to treat preference information as a commitment from the user, rather than the result of a measurement procedure;

[h] Maybe by accepting it, or by providing a direction to their effort in order to obtain a more satisfactory result on the next occasion.

ii) to drop the objective of learning a resolute aggregation procedure, described by a correct value for the preference parameter; and

iii) to clarify the reasoning supporting the elicitation.

1.3.2 A dialectical take to preference information

We make several strong assumptions concerning the preference information collected during the decision aiding process, and fueling the elicitation and explanation processes.

While the process of collecting information needs to be meticulously carried out, with a constant consideration for unwanted behavioral effects, in order to maintain accountability, we will abstract them out of the scope of this report.

Preference information, as any information, is never perfect: the decision maker can be distracted, tired or confused, and report information that does not reflect their actual attitude; worse, they might change their mind during the elicitation process—maybe *because* of the process; inter-human communication is often lossy, and there could be misunderstanding between the decision maker and the analyst. Nevertheless, the counterpart to accountability is to treat the preference information as if it were perfect. We make the normative assumption that the dialog between the decision maker and the analyst is not only an amiable exchange, but that every speech act is a commitment, which is binding w.r.t. the elicitation process and the shape of the aggregation procedure. Of course, we know this assumption is purely virtual, and would eventually be disproved in short order in an actual elicitation process, but its aim is to establish a baseline, by taking the preference information obtained at face value, before considering it as *defeasible*. Therefore, throughout this report, we assume the collected information is free of noise. Meanwhile, we shall consider with great care the fact that the elicitation process occurs under a budget: providing preference information is a demanding process for the decision maker, in terms of time spent and of cognitive burden. Therefore, we assume that preference information might be *incomplete*, and that the aggregation procedure that, ideally, would correspond to the view of the decision maker, is not fully known.

Preference about the model. As the form of the model contributes to the shaping of the preference structure, it ought to be itself tied to some preference information. This point could indeed prove to be crucial, should the assumption concerning the model be challenged by some stakeholder of the decision situation. Throughout this report, we shall adopt a normative stance, and consider the structure of the aggregation model is axiomatic. A

reader interested in breaking down a monolithic axiom, such as 'preferences are additive' or 'preferences are noncompensatory' into finer grains of reasoning could usefully look into the literature concerned with the characterization of models, e.g. [Fishburn, 1976, Krantz et al., 1971, Wakker, 1989, Bouyssou and Marchant, 2007b], as well as the device proposed by [Ouerdane, 2010, Ouerdane, 2009], that identifies and uses properties of a collection of simple models to integrate the question of model selection into the framework of defeasible reasoning.

Holistic preferences. Throughout this report, besides preference information concerning the model, we opt to restrict the input permitting to fit the aggregation procedure to a particular decision situation to *holistic preferences*, under the form of, e.g.:

- 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.

In other words, if we try to elicit an aggregation procedure for a given problem type, the preference information is a *partial function* of this type, that should be extended by the returned aggregation procedure into a total function. There are several advantages to this approach:

- *tangibility*—the information gathered reflects the opinion of the decision maker about alternatives, of which they may have a direct experience, but not about abstract artifacts of a model;
- *unambiguity*—each piece of preference information forms a unit of meaning, regardless of the context. In particular, its interpretation is agnostic to the model. It is also complete, and does not rely on specific assumptions for completion[i];
- *precision*—preference information is described in a bitwise manner, as explicit answers to questions with a binary outcome, rather than infinitely precise equivalence queries as often assumed in the multi-attribute value theory literature [Keeney and Raiffa, 1976, Hammond et al., 1998];
- *universality*—such preference statements can be encountered in many applicative situations, coming from as different an angle as Machine Learning, Decision Analysis or Social Choice. Nevertheless, these settings

[i] As opposed, e.g. to a statement such as 'I prefer blue cars to red cars': does it mean that the speaker prefers *any* car that is blue to *any* car that is red, or merely that, *everything else being equal*, they prefer cars that are blue instead of red?

generally differ concerning the acquisition of preference information: is it given as a whole, picked up one piece at a time, or actively sought for? We discuss some of the bearings of the mode of acquisition of the preference information on the elicitation process in Section 2.4.1.

Additional restrictions. We consider placing constraints on the preference information, either because they allow to represent known facts about the world into our model of the decision aiding process, or because limiting the expressiveness of the decision maker might help in channeling the elicitation process while promoting accountability. For example, in Chapter 3, we consider binary reference scales, that are encountered when the preferences expressed by the decision maker according each point of view reference no more than two levels. Besides luck, such a tight reference set is the consequence of one of these two situations :

- *attributes are themselves binary* : present or absent features, passed or failed checks, etc. Also, such binary attributes may result from any model relying on subset comparisons, such as comparisons between coalition of criteria, or pan-balance comparisons encountered in extensive measurement problems.

- *when expressing preference statements, the decision maker is deliberately restricted to comparing between prototypical alternatives specifically chosen so as to achieve a performance level chosen between two reference values.* This process is supposed to help the decision maker focusing on the main aspects of the preference problems, by limiting the number of changing parts between alternatives, and by referring to carefully chosen reference values, serving as anchors. This technique is used in the field of experimental design (yielding the one-factor-at-a-time or the factorial experiments methods), as well as in multicriteria decision aiding. For instance, the MACBETH method [Bana e Costa and Vansnick, 1995, Bana e Costa et al., 2008] is based on binary alternatives : the decision maker is asked to express preference between prototypical alternatives, traditionally referencing either a *neutral* level (for technological products, representing the attribute of a mid-range, available product), or a *high* level (representing the attribute of a luxury product, or a hypothetical performance demanding a technological breakthrough).

In the same chapter, we also consider the notion of *preference-swaps*, the simplicity of which being evaluated by their *order*. We naturally consider restricting the preference information to low-order preference-swaps.

1.3.3 Robust elicitation as deductive reasoning

Conventional elicitation aims at providing an aggregation procedure, and takes pride in doing this whatever the context—in the case where the aggregation model is overconstrained by the preference information, as well as in the case where it is underconstrained. We propose to drop this goal altogether, and replace it by three components, each one operating according to its own reasoning paradigm:

- a *deductive* component, in charge of drawing conclusions entailed by the statements composing the preference information and the normative stance about decision embodied by the aggregation model;
- a *defeasible* component, in charge of relaxing some of these assumptions, preferably in an accountable way, should the deductive component detect an inconsistency inside the preference information, or an incompatibility between the preference information and the aggregation model;
- an *inductive* component, in charge of making an educated guess in the choice of a specific preference parameter, should the preference information prove consistent, but incomplete—another option to consider would consist in guessing an appropriate *recommendation*, should the deductive component yield an irresolute answer.

In this report, we focus on the deductive component. We connect it to the *robust disaggregation* approach. Instead of returning the most representative model, the *robust* approach draws two sets of conclusions from the preference information, in the form of an *irresolute aggregation procedure*, i.e. aggregation procedures that yield a nonempty set of results, instead of a singleton, of the appropriate type. Given a query, *possible* adjudications are those yielded by at least one aggregation procedure compatible with the preference information. Conversely, *necessary* adjudications are yielded by every aggregation procedure compatible to the preference information, i.e. any other adjudication of the same query is *impossible*.

Ancestry. This approach has a diverse ancestry, and can be traced throughout the history of Operations Research and Artificial Intelligence. In Decision Theory, [Wald, 1950] proposes a non-probabilistic decision making model inspired by *maximin* models of game theory; according to this model, decisions are ranked on the basis of their worst-case outcomes. In Optimization, the so-called *robust approach* [Verdú and Poor, 1984, Ben-Tal et al., 2009] implements Wald's criterion and aims at providing a solution that is good whatever the unknown state of nature is. In Non-Monotonic Logic, the *credulous* and *skeptical* reasoning types allow to deal with unresolved conflicts [Strasser and Antonelli,

2018], and respectively correspond to possible and necessary outcomes; *possibility theory* [Dubois and Prade, 1988] proposes to represent uncertainty through measures of possibility and necessity, and *bipolar reasoning* [Dubois and Prade, 2008] explicitly deals with this dual representation of information. In supervised Machine Learning, [Mitchell, 1982] defines the *version space* as the set of all models compatible with the learning data, partially ordered by set inclusion; the most general model permits to draw possible/optimistic conclusions, while the most specific model draws necessary/pessimistic conclusions. In Multiple Criteria Decision Aiding, the *robust ordinal regression* is formally introduced in [Greco et al., 2008] in the context of the additive value model, implementing notions already sketched in [Hazen, 1986] and [Salo and Hamalainen, 2001], then extended to a number of models ([Angilella et al., 2010, Corrente et al., 2014, Vetschera, 2017]); under the name of *preference inference*, it also appears in [Wilson, 2009, Wilson et al., 2015, Spliet and Tervonen, 2014]. In Decision Theory, [Ok, 2002] considers the representation of decisional behavior with a set of utility functions; [Giarlotta and Greco, 2013] formalizes the abstract framework of necessary and possible preference relations we inscribed the present work in.

1.3.4 Explaining robust adjudications

We would like to provide insight to the decision maker—and an analyst, if they are human—concerning the reasons leading to consider that a given adjudication is impossible, possible, or necessary.

Purposes of an explanation. At this stage, the aggregation procedure should be considered to be in a transient state, under elaboration. We would like to provide tools, not necessarily geared towards the validation of the aggregation procedure—determining its adequacy—but towards its scrutiny. For instance, these explanations could provide, e.g.:

- insight about the *specific adequacy* of a recommendation—ensuring the recommendation results from a correct application of the procedure;
- a *causal assignment* for a recommendation—unveiling its causes, in this case with a transductive[j] flavor, as the retrieved causes shall consist in holistic preference statement, referring to what can be conceived of as prototypical cases;
- a *cognitively friendly*—i.e. easy to follow—path of reasoning towards the given adjudication;

[j] Reasoning from particular to particular—see e.g. [Gammerman et al., 1998a, Pirlot et al., 2016]

- an entry point into the analysis of causes, maybe providing leverage in contesting the causes leading to an unsatisfactory behavior of the aggregation procedure.

Production of an explanation. As to the problem of the production of these explanations, we propose to formalize the robust elicitation as a *decision problem* in the sense of Computer Science—a yes-no question of the input values[k]. Informations that can be deduced from the preference information are, indeed, queries for which a different outcome could not have been represented in the model, together with this preference information. Therefore, we define the *inverse problem* of a model, as the question 'can this aggregation procedure be represented in this model?'. We note that this problem is closer to the question of the representation of a model, central in Conjoint Measurement, than the mainstream XAI trend interested in reasoning on the inverse problem *of the aggregation procedure*—e.g., under which conditions this procedure would have yield a better result?

The inverse problem can be used to provide *explanations*, based on the *certificates* of a positive instance (for supporting a possible adjudication) or a negative instance (for supporting impossible or necessary adjudications).

The workflow corresponding to our proposed framework is presented in Table 1.2

Assessment of an explanation. The purposes of an explanation define long-term goals. In the scope of this report, we would rather focus on the production of an explanation.

Research question 1 (computation)**.** *How difficult is it to produce an explanation?*

We expect this question to require notions and tools from the field of Computational Complexity.

Although they are of a formal nature, the explanations produced should eventually be presented to humans.

Research question 2 (simplicity)**.** *Can we keep the explanations simple enough?*

[k]Not to be confounded with the topic discussed in e.g. [Colorni and Tsoukiàs, 2013]. In the context of decision aiding, the words 'decision', 'problem' and 'model' are dangerously overloaded with different but connected meanings.

Step	Action
A	Collect holistic preference information, under the form of a set of queries, together with their adjudication. This knowledge base functions as a learning set for the problem of fitting the preference parameter.
B	Transcribe each piece of preference information into a set of constraints bearing on the value of the preference parameter.
C	Adjudicate new queries, not belonging to the learning set, by considering all the aggregation procedures corresponding to the chosen model and the compatible parameters (i.e. satisfying the constraints). Predicates satisfied by every possible adjudications are *necessary*. Predicates satisfied by no possible adjudication are *impossible*.
D	Support the adjudication with an explanation.

Table 1.2: The *robust disaggregation framework* for accountable adjudication. *Adjudication* can be e.g. the assignment of an alternative to a category (for sorting problems) or the specification of pairwise preference between two alternatives (for problems based on comparisons, e.g. choice or ranking).

Neither natural language generation, nor in vivo experimentation belong to the scope of this report, so the complexity of explanations shall be assessed through proxies, such as length, or number of premises.

Explanation, in general, is based on reformulation. As we strive to give explanations of a deductive process, a natural pair of questions concerns the relationship between the explained system—the robust elicitation process—and the explaining system.

Research question 3 (soundness)**.** *Could we explain 'false' results, claiming the impossibility of an event that could happen or the possibility of an event that cannot happen?*

Research question 4 (completeness)**.** *Can we explain every 'true' result, that can be deducted from the preference information and the model?*

As usual when dealing with multiobjective problems, the separation of these issues is somewhat illusory. An explanation system could be made incomplete by confining it to providing simple explanations, for instance.

1.3.5 The shape of things to come

Besides this introductory chapter, the conclusive Chapter 8 and the appendices, this report is divided in two parts, each devoted to a specific problem type and aggregation model:

- Part I, composed of chapters 2 and 3, considers the problem of comparing pairs of alternatives with an additive value model. Therefore, we are trying to sculpt a binary aggregation procedure, that, when fully elicited, should *adjudicate* the preference between any two alternatives.
- Part II, composed of chapters 4 to 7, deals with the problem of sorting alternatives into categories ordered by level of requirement, with a noncompensatory model. These models consider putting a cardinality constraint on the scales permitting to represent the preference according each point of view, so as to limit trade-offs.

The models chosen are relatively simple and are generally considered 'interpretable', and thus are legitimate candidates when considering to make decision aiding accountable again.

For these specific problem types and aggregation model, we either build upon the existing robust elicitation structure, to design the deductive component:

- the additive value model, assumed in Part I, can be considered as the 'flagship' of aggregation models in Decision Aiding and Social Choice and still serves as a baseline in Machine Learning. It benefits from abundant studies, and in particular, its robust elicitation has been developed in [Greco et al., 2008], where the problem of adjudicating necessary queries is formulated as a *linear program*. We therefore build upon this base, and propose a dual formulation of this problem, with arguably better properties concerning accountability (Chapter 2). Then, we propose to interpret a Farkas certificate, either to directly provide deductive explanations based on the cancellative properties of the model, or to ease the breaking down of a necessary preference statements in elementary parts, in the manner of the even-swaps active elicitation method proposed by [Hammond et al., 1998] (Chapter 3).
- the noncompensatory sorting model, assumed in Part II, benefits from less attention than the additive value model, and has a reputation of computational quagmire when it comes to disaggregation. We give a formal description of the model, accounting for part of its numerous variants (Chapter 4). We formalize the corresponding inverse problem in the language of Boolean satisfiability, prove its NP-hardness, even in the case where there are only two categories (Chapter 5), and obtain results

at least an order of magnitude faster than preceding attempts based on mixed integer programs (Chapter 6). We propose explanations based on unsatisfiable subsets of clauses of negative instances, in the manner of [Junker, 2004, Besnard et al., 2010], as well as practical ways to use them in decision aiding situations inspired by recent concerns coming e.g. from the legal domain (Chapter 7).

Finally, Chapter 8 wraps up this report, by putting the contributions into perspective.

1.4 Formalization

In this section, we detail the assumptions made throughout this report, as well as the notations and writing conventions we use.

1.4.1 Points of view and alternatives

We denote $\mathbb{X} = \{x, y, z, \dots\}$ the set of *alternatives* and $N = \{i, i', \dots\}$ the set of *points of view*. Together, they define the objects of the elicitation dialog. We assume the points of view provide a sense of the relative fitness of alternatives, for which we consider two equivalent representations:

- *preference profiles*, a tuple $\langle \succsim_i \rangle_{i \in N} \in (\mathbb{X} \times \mathbb{X})^N$ of *total preorders* over alternatives—binary relations that are transitive and connected. This representation is often used in Social Choice or when representing preferences with an outranking relation. Table 1.3 provides an illustration with a situation detailed in Section 7.2 where each point of view corresponds to the views of a juror in a jury $N = \{\text{👩}^1, \text{👩}^2, \text{👩}^3, \text{👩}^4, \text{👩}^5\}$ gathered to assess the fitness of a number of candidates $\{a, b, c, d, e, f\} \subset \mathbb{X}$.
- *performance tables*, where an alternative $x \in \mathbb{X}$ is described by a tuple of performance scalars $\langle x_i \rangle_{i \in N}$ encoding their fitness according to each point of view $i \in N$ on an ordinal scale $(K_i, \geq_i)$. Tables 1.4 and 1.5 provide two illustrations with alternatives representing, respectively, hotels, a situation considered throughout Part I, and cars, taken from Chapter 4.

The two representation are tied by the tuple of *isotonic homomorphisms*[l] $\langle (\mathbb{X}, \succsim_i) \to (K_i, \geq_i), x \mapsto x_i \rangle_{i \in N}$, usually called *criteria*. Such a representation is possible as long as the cardinality of the scales $|K_i|$ allows to accommodate for the equivalence classes of the relations $\sim_i$[m].

[l] *Isotonic homomorphisms* preserve the order structure. Unsophisticated people call them *increasing functions*.

[m] When a scale is continuous, representation is possible when $\mathbb{X}/\sim_i$ has a denumerable order-dense subset—$B \subset A$ is *order dense* in $(A, \succ)$ when $\forall a, b \in A,\ a \succ b \Rightarrow \exists c \in B :\ a \succ c \succ b$.

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

Table 1.3: A *preference profile*, detailing the ordinal preferences of jurors over candidates. Here each profile is a total order—there are no ties.

Hotel	Comfort	Restaurant	Commute time	Cost
h_A	$4^\star$	no	35 min	120 $
h_B	$4^\star$	yes	50 min	160 $
h_C	$2^\star$	yes	20 min	50 $
h_D	$2^\star$	no	30 min	40 $

Table 1.4: A *performance table*: alternatives are hotels, assessed according to the point of view of comfort, the presence of a restaurant, the commute time needed to reach the city center, and the cost per night.

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 1.5: A *performance table*: alternatives 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).

In both representation, the notions of alternatives and points of view are bound together so that:

i) for all purposes, the points of view, taken together, give a full account of an alternative, i.e. alternatives that would be considered as similar according to all points of view are indiscernible;

ii) for the sake of discussion, we allow for any configuration of preferences[n].

Therefore, we assume that every preference profile is acceptable, or equivalently that alternatives can achieve any level of performance, i.e. $\mathbb{X} \equiv \prod_{i \in N} K_i$.

Dominance occurs when the points of view unanimously prefer an alternative over another, defining a preorder $\mathcal{D} \subset \mathbb{X}^2$:

$$\mathcal{D} := \bigcap_{i \in N} \succsim_i . \tag{1.1}$$

1.4.2 Problem type

This report focuses on two different problem types: comparing pairs of alternatives, and sorting alternatives into predefined categories ordered by level of requirement. This is an opportunity to define, in the abstract, a *problem type* T as a functional set, containing the potential aggregation procedures. T is described as a specific subset of the set of functional relations from a domain T^{in} containing the queries to a codomain T^{out} containing the adjudication.

a) *Sorting:* there is a set of categories ordered by level of requirement $\{C^1 \prec \cdots \prec C^p\}$, so that T_{sorting} is the set of isotonic homomorphisms mapping each query in $T^{in}_{\text{sorting}} := (\mathbb{X}, \mathcal{D})$ to an adjudication in $T^{out}_{\text{sorting}} := \{C^1 \prec \cdots \prec C^p\}$; elements ϕ of T_{sorting} are unary functions, yielding an *absolute* judgment on an alternative on its own, with the following semantic (e.g. assuming $p = 3$):

$$\forall x \in \mathbb{X},\ \phi(x) = \begin{cases} C^1, & \text{if } x \text{ is deemed } \textit{bad}; \\ C^2, & \text{if } x \text{ is deemed } \textit{average} \\ C^3, & \text{if } x \text{ is deemed } \textit{good}. \end{cases}$$

b) *Pairwise comparison:* T_{pairwise} is the set of functions ϕ from $T^{in}_{\text{pairwise}} := \mathbb{X}^2$ to $T^{out}_{\text{pairwise}} := \{\text{YES}, \text{NO}\}$, *compatible with* (or *extending*) dominance: $\phi(\mathcal{D}) = \{\text{YES}\}$; ϕ is a *binary* function yielding a comparative judgment between two alternatives, so that the binary relation $\succsim := \phi^{-1}(\{\text{YES}\})$ has the semantics of an *outranking relation* [Roy, 1991]:

- if $x \succsim y$ but $y \not\succsim x$, then x is *strictly preferred* to y;
- if $x \not\succsim y$ and $y \succsim x$, then y is *strictly preferred* to x;
- if both $x \succsim y$ and $y \succsim x$, then x and y are *indifferent* or *equally preferred*;
- if both $x \not\succsim y$ and $y \not\succsim x$, then x and y are *incomparable*.

[n] Even extravagant ones, such as extremely cheap luxury hotels without a restaurant—one is allowed to dream.

c) *Choice:* T_{choice} is the set of *choice functions* that, given an query consisting of a tuple of alternatives, yield the fittest alternative among those in the input tuple;

d) *Ranking:* T_{ranking} is the set of functions that, given a query consisting of a tuple of alternatives, yield a permutation of this tuple sorted by increasing overall fitness.

1.4.3 Aggregation models, the inverse problem and necessary adjudications

Given a problem type T, e.g. pairwise comparison in Part I, or sorting in Part II, we define a *model of type* T as a pair $(\Omega, \langle\phi_\omega\rangle_{\omega\in\Omega} \in T^\Omega)$, where Ω is the *parameter space* and $\langle\phi_\omega\rangle_{\omega\in\Omega}$ is a family of aggregation procedures with type T indexed by a *preference parameter* $\omega \in \Omega$.

Definition 1.1 (inverse problem)**.** *Given a problem type T and a model $\mathcal{M} \equiv (\Omega, \langle\phi_\omega\rangle_{\omega\in\Omega} \in T^\Omega)$ of type T , the inverse problem Inv-$\mathcal{M}$ is the decision problem consisting in, given an instance $\mathcal{P}$ where $\mathcal{P}$ is a function of type T, to decide whether there is a value of the preference parameter $\omega^\star \in \Omega$ such that $\phi_{\omega^\star} \supset \mathcal{P}$. If $\mathcal{P}$ is a negative instance, we say it is inconsistent with $\mathcal{M}$, or that it cannot be represented in $\mathcal{M}$.*

The condition $\phi_{\omega^\star} \supset \mathcal{P}$ states that the aggregation procedure $\phi_{\omega^\star}$ extends the preference information, or, more precisely, for each preference statement $(q, a) \in \mathcal{P}$, the query q is adjudicated in the same manner by the aggregation procedure $\phi_{\omega^\star}$, i.e. $\phi_{\omega^\star}(q) = a$.

We denote $\Omega_{\mathcal{P}}^{\mathcal{M}}$, or, when the model is made clear by the context, simply $\Omega_{\mathcal{P}}$, the set of preference parameters *compatible* with the preference information $\mathcal{P}$:

$$\Omega_{\mathcal{P}}^{\mathcal{M}} := \{\omega \in \Omega : \ \phi_{\omega^\star} \supset \mathcal{P}\}. \tag{1.2}$$

At this stage of the elicitation process, we identify three different situations, according to the size of the set $\Omega_{\mathcal{P}}^{\mathcal{M}}$:

- either $\Omega_{\mathcal{P}}^{\mathcal{M}} = \varnothing$—the preference information is *inconsistent* with the model, and the situations should be considered through the prism of *defeasible reasoning*;
- or $|\Omega_{\mathcal{P}}^{\mathcal{M}}| > 1$—the preference information is *incomplete* and there is some ambiguousness w.r.t. the value of the preference parameter;
- finally, for the sake of exhaustiveness, when $\Omega_{\mathcal{P}}^{\mathcal{M}}$ is a singleton, elicitation is complete.

In this report, we assume that the preference information is never inconsistent with the model, i.e. $\Omega_{\mathcal{P}}^{\mathcal{M}} \neq \emptyset$.

Definition 1.2. *Given a problem type T, a model $\mathcal{M} \equiv (\Omega, \langle \phi_\omega \rangle_{\omega \in \Omega} \in T^\Omega)$ of type T, some preference information $\mathcal{P}$ of type T consistent with $\mathcal{M}$ and a statement $\sigma \in T$, $\sigma \equiv (q, a)$, we say:*

- *σ is* possible *if $\mathcal{P} \cup \{\sigma\}$ is a positive instance of Inv-$\mathcal{M}$;*
- *σ is necessary if $\forall \sigma' \in \{(q', a') \in T : q' = q \text{ and } a' \neq a\}$, $\mathcal{P} \cup \{\sigma'\}$ is a negative instance of Inv-$\mathcal{M}$.*

Part I

Comparing with an additive model

Foreword

This part presents original contributions to the scientific knowledge about the necessary preference relation in the case where preferences are described by an *additive value model*. This model is simple and elegant, yet plays a central role in Decision Aiding, Social Choice and Machine Learning alike.

We give a brief definition of the comparison of alternatives using an additive value model, highlight the most preeminent qualities of this approach, and give a quick glimpse of domains related to decision making where this model serves as a baseline. We list the research questions addressed in this work, give an overview of the contents of each chapter, and list our published papers that serve as a basis for this text.

Comparing alternatives with additive values

A preference relation $\succsim$ follows a *value model* when the overall desirability of an alternative can be measured by a numerical score, the higher, the better. Technically, there is a numeric function V mapping alternatives to real numbers:

$$V : \mathbb{X} \to \mathbb{R} \tag{I.3}$$

Scores are then compared to derive preferences:

$$\forall x, y \in \mathbb{X},\ x \succsim y \iff V(x) \geq V(y) \tag{I.4}$$

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[a].

The very general framework offered by value models does not say much about the manner in which the various points of view are aggregated. In order to extend the dominance relation $\mathcal{D}$, the function V is only required to be nondecreasing according to each point of view.

Throughout this part, we assume that preference proceeds from an *additive value* model, that postulates a specific, separated form for the value V:

$$\exists \langle v_i \rangle \in \prod_{i \in N} (\mathbb{X} \to \mathbb{R}) \; s.t. \; \forall x \in \mathbb{X}, \; V(x) = \sum_{i \in N} v_i(x) \tag{I.5}$$

In order for the preference relation to extend dominance (see Equation (1.1)), all marginal value functions need to reflect preference according to the point of view it represents:

$$\forall i \in N, \; v_i \text{ is a nondecreasing function from } (\mathbb{X}, \succsim_i) \text{ to } (\mathbb{R}, \geq). \tag{I.6}$$

Interestingly, the additive value model, which is by design an 'aggregate then compare' model, can also be expressed in the 'compare then aggregate' paradigm as well (see Figure 1.3). Each point of view $i \in N$ compares alternatives $x, y \in \mathbb{X}$ through the function c_i such that $c_i(x, y) = v_i(x) - v_i(y)$. These orientations are then aggregated with a simple sum. Finally, the sign of the result is mapped to preference: $\{(+, x \succ y), (0, x \sim y), (-, x \prec y)\}$.

The linear canvass of additive values is quite an asset when considering indirect elicitation (see Section 1.2.3). The disaggregation of preference statements into constraints on the preference parameter naturally yields a linear formulation. Therefore, the problem of optimizing a loss function to induce a model reduces to a linear [Siskos et al., 2005, Greco et al., 2008] or semidefinite program [Bous et al., 2010]—depending on the choice of the loss function—that can be solved efficiently.

Usage of the additive value model

Multiple criteria decision aiding. In *multiple criteria decision aiding* (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:

[a]There is, actually, a third condition called *order density*. It concerns the number of equivalence classes of the symmetric part of the relation, and prevents, for example, that it is so large that as to exceed the capacity of the set of real numbers. This theoretical consideration has no practical bearing whatsoever on the aggregation of preferences in a decision aiding situation.

- it connects the field with conjoint measurement and mathematical psychology [Fishburn, 1967a, Krantz et al., 1971];
- it serves as the basis of very popular methods, such as the *multi-attribute value theory* (MAVT) [Keeney and Raiffa, 1976], the *analytical hierarchical process*[b] (AHP) [Saaty, 1990], or MACBETH [Bana e Costa and Vansnick, 1994];
- it served as a platform for the major innovations of the domain:
  - the indirect elicitation via standard sequences [Keeney and Raiffa, 1976];
  - the indirect elicitation via disaggregation [Siskos et al., 2005];
  - active learning via the *even-swaps* method [Hammond et al., 1998];
  - the robust disaggregation [Greco et al., 2008].

For many reasons, and above all simplicity, the additive model was, and still is, a model of choice in other fields interested in aggregating evaluations.

Naive procedures. Evaluation of performance is ubiquitous, and so is the need to aggregate evaluations assessed from diverse points of view. People do this everywhere, all the time, without a specific background in Decision Theory[c]. It may be highly suspected that the method of choice for the layman is the *weighted sum*—which is a particular case of the additive value model, where the marginal values are linear—as it is easy to compute (especially with a spreadsheet) and it offers a modicum of control through the selection of weights.

Conjoint measurement. Conjoint measurement is interested in situations where several latent attributes combine into an observable one, and tries to obtain conditions favorable to a numeric evaluation of both the observable and the latent quantities. The way the latent attributes combine into the aggregated one is a central assumption in this approach, and the additive value model, with two attributes, was the one permitting the fundamental breakthroughs of [Debreu, 1960] and [Luce and Tukey, 1964]. From this initial point, more sophisticated models were considered: additive values with any number of attributes, polynomial values, non-compensatory models with discrete levels, models permitting to account for interactions between inputs. Meanwhile, the conditions in which the additive value model can legitimately be assumed as

[b] The AHP is described in a multiplicative value paradigm, which is obviously equivalent to the additive value model.

[c] Gasp!

the latent aggregator have been thoroughly investigated by researchers in the field of mathematical psychology, e.g. [Miller, 2018].

Social choice. The aggregation of preference has been the chief preoccupation of social choice since its inception at the end of the 18th century with de Borda and Condorcet.

The basic setting of *voting* is squarely the same than the one discussed in this report: preferences according to each point of view are given (as opposed to conjoint measurement) and need to be aggregated via a *voting rule*. In the context of real-life applications of voting, the points of view can range in the millions, which calls for rules that are computationally very efficient. Among these, the *positional scoring rules* map the ordinal preferences expressed by each point of view, supposed to be complete orders, to scores, which are then summed up across all points of view: they enforce the additive value model. Some popular voting rules—e.g. Plurality, Borda, k-Approval, see [Brandt et al., 2016]—are positional scoring rules, while others—e.g. Single Transferable Vote, Copeland, Kameny—are not.

In *combinatorial auctions*, agents bid on bundles of objects. Their behavior can be modeled, with a representation of their preference relation between subsets of object. A (very) simplifying assumption is to neglect interaction between objects—either positive *complementarity* or negative *substitutability*, see e.g. [Brandt et al., 2016]—and use an additive value model, where the value of a bundle is simply the sum of the values of the objects composing the bundle.

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 kind of desirability, this behavior can be considered through the prism of the aggregation of evaluations stemming from multiple points of view. Linear classifiers, producing boundaries between categories that are affine hyperplanes, are still widely used, even though they are often outmatched in terms of accuracy by more recent techniques, such as random forests or neural networks (see e.g. [Bishop, 2006]). Nevertheless, as of 2018, advances in machine learning still opt, more often than not, to present their results as upgrades from the baseline offered by the additive model.

Research questions

There have been numerous contributions to and applications of the additive value model. Most of them focus on modeling issues—i.e. to what extent this

model can be used to describe real situations, or to prescribe solutions—and elicitation issues—how to fit the parameters governing the model in order to capture the phenomenon that needs to be modeled, especially in the presence of uncertain data. An issue that has been less investigated (for notable exceptions, see e.g. [Labreuche, 2011] and [Cailloux and Endriss, 2014]), and remains difficult [Procaccia, 2018], is the question of the accountability of recommendations based on an induced model.

Our contributions detailed in this part try to address this issue. Until now, elicitation has mostly been considered through an Operations Research lens, as a problem needing solving. We propose to consider it also through an Artificial Intelligence lens, as a *reasoning* needing to be formalized. In such a KR[d] framework, the recommendation (necessary preference) is the *product* of this reasoning, and explanation could be a *certificate* of this process.

Research question I.1 (Formalizing the reasoning?)**.** *Could the notions and tools put forward for the 'robust elicitation' of the additive value model benefit from being streamlined by borrowing notions—such as those of 'knowledge base' and 'inference engine'—from the community interested in the representation of knowledge and reasoning?*

Research question I.2 (From certificate to explanation?)**.** *Under the assumption of additive preferences, what certificate of necessary preference can we devise for a pair of alternatives? How can we compute one? How can we leverage one into providing* explanations*?*

Research question I.3 (Qualifying the explanations?)**.** *How the proposed explanation engines address the challenges identified in the introduction, e.g.* completeness, soundness, simplicity *and* computational *aspects?*

Chapters

Two chapters compose this part.

- Chapter 2 addresses the question I.1 by proposing an encoding of the pairwise preference statements of the preference information, as well as the preference queries that require adjudication. This static encoding permits to formalize the adjudication of queries as an inference problem, and to propose a resolution procedure based on Farkas' certificates, partly answering Question I.2. This procedure greatly improves the interpretability of the adjudication process, without compromising its computational efficiency.

[d]Principles of knowledge representation and reasoning.

- Chapter 3 addresses the second part of Question I.2, by proposing to build explanations engines upon the Farkas' certificates of necessary preference, with integral coefficients. Two approaches to explanation are investigated: one leverages the cancellative properties of the additive value model; the other is reminiscent of the *even-swaps* method [Hammond et al., 1998], and proposes to break down a preference statement into a sequence of transitive, simple preference statements. These approaches are discussed in the light of Question I.3.

Featured contributions

The encoding and its properties, the characterization of necessary preference with Farkas' certificates, and the explanation engine based on sequences of preference swaps, are introduced in [Belahcene et al., 2017a]. This work has benefited from many fruitful interactions with researchers interested in multiple criteria aiding, and has been greatly improved since.

2

Robust elicitation as reasoning

2.1 Introduction

This chapter is devoted to fundamental knowledge representation and reasoning aspects, and introduces a novel encoding of preferences for step B, of the robust disaggregation framework (see Figure 1.2) , and a novel process for adjudicating a query for step C that, together, hopefully help improve the interpretability of the model. The latter deals with step D builds upon these formal results by proposing two ways of explaining necessary preference statements, discusses their respective merits and paves the way to mixing and matching them into a joint approach.

2.1.1 State of the art

The problem of determining, given

i. some preference information under the form of a tuple of ordinal, pairwise holistic preference statements;

ii. a pair of alternatives $x, y \in \mathbb{X}$;

whether x is either *necessarily*, *possibly* or *impossibly* preferred to y under the assumption that the preference structure can be described in the additive value model, has been addressed by [Greco et al., 2008]. This paper introduces two definitions formalizing the fundamentals of the necessary preference relation, then proposes a characterization with a linear program[a].

[a] A linear program is an optimization problem of the form $\min\{f(x), x \in Q\}$, where f is a linear function and Q is a polytope in $\mathbb{R}^n$. It can be solved in polynomial time [Karmarkar,

Definition 2.1 (Additive value functions compatible with the preference information)**.** *Given some preference information $\mathcal{P} \subset \mathbb{X} \times \mathbb{X}$, a tuple of functions $\langle v_i \in \mathbb{X} \to \mathbb{R} \rangle$ is* additively compatible with the preference information $\mathcal{P}$ *if :*

(i) for each point of view $i \in N$, v_i is a nondecreasing function from $(\mathbb{X}, \succsim_i)$ to $(\mathbb{R}, \geq)$;

and :

(ii) for each pair of alternatives $(a, b) \in \mathcal{P}$, $\sum_{i \in N} v_i(a) \geq \sum_{i \in N} v_i(b)$.

The first condition ensures compatibility with *dominance* $\mathcal{D}$, the second enforces compatibility with the *preference information* $\mathcal{P}$. For the sake of simplicity, we restrict ourselves to the case where the preference information only contains statements of weak preference, with the semantic 'at least as good as'. Consequently, whatever the preference information, the set of tuples of value functions additively compatible with it is never empty, because any tuple of constant values satisfies the constraints. Also, *any* preference is possible, and we restrict the investigation to the necessary preference relation. The extension to strict preference statements is straightforward, but cumbersome.

Definition 2.2 (Necessary preference under the assumption of additive utility)**.** *[Greco et al., 2008]. Given two alternatives $x, y \in \mathbb{X}$ and some preference information $\mathcal{P} \subset \mathbb{X} \times \mathbb{X}$, we say x is* necessarily preferred *to y and we note $(x, y) \in \mathcal{N}_\mathcal{P}$ if the inequality $\sum_{i \in N} v_i(x) \geq \sum_{i \in N} v_i(y)$ holds for every tuple of functions $\langle v_i \in \mathbb{X} \to \mathbb{R} \rangle$ additively compatible with the preference information $\mathcal{P}$.*

Obviously, any preference relation stemming from an additive value model and compatible with the preference information $\mathcal{P}$ extends $\mathcal{N}_\mathcal{P}$, and $\mathcal{N}_\mathcal{P}$ is exactly the intersection of all such preference relations.

Also, $\mathcal{N}_\mathcal{P}$ is reflexive, transitive and extends both $\mathcal{D}$and $\mathcal{P}$.

Inference via primal feasibility of a linear program Definition 2.2, introduced by [Greco et al., 2008], can be directly leveraged to formulate a linear program (LP) permitting to decide, for any pair of alternatives $x, y \in \mathbb{X}$, whether x is necessarily preferred to y or not.

Proposition 2.1. *[Greco et al., 2008]. Given some preference information $\mathcal{P} \subset \mathbb{X} \times \mathbb{X}$ and two alternatives $x, y \in \mathbb{X}$, $(x, y) \in \mathcal{N}_\mathcal{P}$ if, and only if, the linear program* $\min \left(\sum_{i \in N} v_i(x) - \sum_{i \in N} v_i(y), \langle v_i(z) \rangle \in Q \right)$ *has a nonnegative solution,*

1984].

where Q is the polytope defined by the constraints:

$$\begin{cases} v_i(z) \in \mathbb{R} & \forall i \in N,\ \forall z \in \{x\} \cup \{y\} \cup \bigcup\limits_{(a,b)\in\mathcal{P}} \{a,b\} \\ \sum\limits_{i\in N} v_i(a) \geq \sum\limits_{i\in N} v_i(b) & \forall (a,b) \in \mathcal{P} \\ v_i(c) \geq v_i(d) & \forall i \in N, (c,d) \in \mathcal{R}_i \end{cases}$$

Any pair of alternatives $(x, y) \in \mathbb{X}^2$ partitions the parameter space into two half-spaces separated by a hyperplane $\sum_{i\in N}(v_i(x) - v_i(y)) = 0$, one where x is preferred and one where y is preferred. Thus, the parameters compatible with the preference information reside in a polytope of the parameter space. A query is adjudicated in a necessary manner if, and only if, this polytope is entirely on one side or another of the hyperplan corresponding to the query, or ambivalent when the boundary hyperplane intersects the polytope.

While effective, this approach is not fully satisfactory. As a direct reformulation of the definition of the relation $\mathcal{N}_\mathcal{P}$, it does not offer any perspective on the issue. Moreover, it inscribes the adjudication problem into an optimization framework that can be considered inscrutable by the layman — here, meaning anyone who is not well-versed in the arcane of Operations research. The endeavor of the entire Part I of this report is to lower the barrier of entry to this problem.

2.1.2 A grim situation

One of the challenges in dealing with the notion of necessary preference is that it formally requires a level of due diligence that seems out of reach: in order to prove that Alice is necessarily preferred to Bob, one has to certify that, of the uncountably many additive values that are compatible with the preference information, there is not a single one that puts Bob ahead of Alice. How can one be sure they have checked them all? In their foundational article [Greco et al., 2008], Greco et al. delegated the superhuman task of checking such claim to a linear programming solver. This creates an unfortunate situation where an analyst and/or a decision maker may feel powerless, deprived of their ability to critically challenge one of the key steps of the decision-aiding process, as they have to surrender their limited mathematical skills to the might of the device equipped with an optimization engine.

The situation is grim. In order to circumvent this deleterious scenario [Spliet and Tervonen, 2014] set out to describe the inference of necessary preference relations in what should be called a rule-based framework, but, in spite of early promising advances (see Sections 2.3.2, Component 4 and 2.4.1),

their results are really weak[b].

2.1.3 A motivating example

The following example could provide a glimmer of hope.

Example 2.1. You need to chose a hotel for a business trip, and you are undecided between four options h_A, h_B, h_C and h_D, described by the performance table 2.1. Such options are evaluated according to four *criteria* :

- the *room comfort*, ranging from $*$ (low) to $*****$ (high);
- the presence of a *restaurant* on the premise, with *yes* preferred to *no*;
- the *commute time* to the convention center, the lower the better;
- the *cost*, the lower the better.

Criteria	Comfort	Restaurant	Commute time	Cost
h_A	$4^\star$	no	35 min	120 $
h_B	$4^\star$	yes	50 min	160 $
h_C	$2^\star$	yes	20 min	50 $
h_D	$2^\star$	no	30 min	40 $

Table 2.1: Performance table of the hotels.

We assume the preference information contains the following statements: $\mathcal{P} \supseteq \{\pi_1, \pi_2, \pi_3\}$, with:
π_1 :=((4* , no , 15 min , 180 $) , (2* , yes , 45 min , 50 $))
π_2 :=((4* , no , 45 min , 50 $) , (4* , yes , 15 min , 100 $))
π_3 :=((2* , yes , 15 min , 180 $) , (4* , no , 30 min , 180 $))

For any additive values V that extends the preference information, the following equalities stand:

- from ((4* , no , 15 min , 180 $) , (2* , yes , 45 min , 50 $)) $\in \mathcal{P}$ we derive:
  $v_*(4^*) + v_r(\text{no}) + v_t(15\text{ min}) + v_\$(180\ \$) \geq v_*(2^*) + v_r(\text{yes}) + v_t(45\text{ min}) + v_\$(50\ \$)$
- from ((4* , no , 45 min , 50 $) , (4* , yes , 15 min , 100 $)) $\in \mathcal{P}$ we derive:
  $v_*(4^*) + v_r(\text{no}) + v_t(45\text{ min}) + v_\$(50\ \$) \geq v_*(4^*) + v_r(\text{yes}) + v_t(15\text{ min}) + v_\$(100\ \$)$
- from dominance for the criterion *restaurant* we derive :

[b] Up to the point where they devote a considerable energy to prove that, finally, necessary preference is not that useful because its so-called 'probability of occurrence' is low.

$$v_r(\text{yes}) \geq v_r(\text{no})$$

Adding these three inequalities leads to:

$v_*(4^*)+v_*(4^*)+v_r(\text{no})+v_r(\text{no})+v_r(\text{yes})+v_t(15\text{ min})+v_t(45\text{ min})+v_\$(180\ \$)+v_\$(50\ \$) \geq v_*(2^*)+v_*(4^*)+v_r(\text{yes})+v_r(\text{yes})+v_r(\text{no})+v_t(45\text{ min})+v_t(15\text{ min})+v_\$(50\ \$)+v_\$(100\ \$)$.

Cancelling terms $v_*(4^*), v_r(\text{no}), v_r(\text{yes}), v_t(15\text{ min})$ and $v_\$(50\ \$)$ appearing on both sides leads to:

$v_*(4^*) + v_r(\text{no}) + v_t(45\text{ min})) + v_\$(180\ \$) \geq v_*(2^*) + v_r(\text{yes}) + v_t(45\text{ min})) + v_\$(100\ \$)$

As this inequality holds for all values compatible with the preference information, it follows that alternative (4* , no , 45 min , 180 $) is necessarily preferred to alternative (2* , yes , 45 min , 100 $).

In Example 2.1, necessary preference is certified in a constructive manner. Rather than following the definition and exhausting the set of preference parameters, it exhibits an algebraic relationship between premises— linear inequalities encoding pieces of preference information— and the desired conclusion. This fortunate scenario is encouraging, but one can wonder about its generality. Is it limited to some lucky cases, or is it a property universally shared by necessary preference statement?

2.1.4 To the rescue: Farkas' lemma

Theory can help here to address this question. Suppose that, indeed, Alice is preferred to Bob under every possible assumption. What we are trying to prove can be framed as the impossibility of finding a value function that simultaneously restores the preference information and puts Bob ahead of Alice. What we need, then, is a formal certificate of infeasibility for this problem. As we have already identified this problem as a linear programming one, we might consider the following 19th century result obtained by Giulya Farkas:

Proposition 2.2 (Farkas' lemma)**.** *Given $E \equiv \mathbb{R}^n$, a vector space of finite dimension over the field of real numbers and $\langle h, f_1, \ldots, f_k \rangle$, a tuple of linear forms on E:*

$\{y \in E \mid h(y) > 0, f_1(y) \geq 0, f_2(y) \geq 0, \ldots, f_k(y) \geq 0\} = \varnothing$

if, and only if,

$(-h)$ *is a linear combination with nonnegative coefficients of $f_1, \ldots, f_k$.*

Farkas' lemma is a solvability theorem for a finite system of linear inequalities. It is a powerful result, foundational of the strong duality theorem for linear optimization, and Karush, Kuhn and Tucker conditions for non-linear

optimization. Moreover, it is closely related to the 'lucky observation' made in example 2.1:

- The observation made in the example corresponds to the *obvious* orientation of the Farkas equivalence—when a linear form h is a linear combination with nonnegative coefficients of the linear forms $f_1, \ldots, f_k$, it is nonnegative whenever the forms $f_1, \ldots, f_k$ are all nonnegative. The fact that this property derives from basic, highschool-level algebra is a blessing from the points of view of accountability, interpretability, and explainability.
- On the other hand, the strong result obtained by Farkas lies in the universality of this easy case. It means that the observation made in Example 2.1 is actually not due to luck but a manifestation of a deeper phenomenon. The form of a generic, universal *certificate of infeasibility* is a tuple $\langle \lambda_1, \ldots, \lambda_k \rangle$ of nonnegative coefficients such that $-h = \sum_{i=1}^{k} \lambda_i f_i$. Interestingly, framing the decision problem 'is Alice necessarily preferred to Bob' as the search for the coefficients of a possible certificate is also a linear optimization problem, which is the dual problem of the original formulation.

2.1.5 Towards an interpretable framework for necessary preference

We note that Proposition 2.1, which is the state-of-the-art definition for necessary preference under the assumption of additive preferences, is almost suitable for processing by Farkas' lemma. We recast the requirement of nonnegativity of the linear program as the infeasibility of satisfying simultaneously all the constraints defining the polytope Q as well as $\sum_{i \in N} v_i(y) > \sum_{i \in N} v_i(x)$. We then move every variables into the RHSs, in order to write the problem in canonical form, i.e. where the feasible region is a polytope resulting from the intersection of a tuple of half-spaces:

Lemma 2.3 (Canonical representation of $\mathcal{N}_{\mathcal{P}}$)**.**

$$(x, y) \in \mathcal{N}_{\mathcal{P}} \iff \textit{there is no vector } \langle v_i(z) \rangle \textit{ for } i \in N \textit{ and } z \in \{x\} \cup \{y\} \cup \bigcup_{(a,b) \in \mathcal{P}} \{a, b\} \textit{ such that:}$$

$$\begin{cases} \sum_{i \in N} v_i(y) - \sum_{i \in N} v_i(x) < 0 & \\ \sum_{i \in N} v_i(a) - \sum_{i \in N} v_i(b) \geq 0 & \forall (a, b) \in \mathcal{P} \\ v_i(c) - v_i(d) \geq 0 & \forall i \in N, (c, d) \in \succsim_i \end{cases}$$

The remainder of this chapter is devoted to the introduction of a streamlined version of the idea of working with certificates of infeasibility. It is organized as follows: Section 2.2 provides the nuts and bolts of the representation, and should be viewed as a repository of tools; Section 2.3 illustrates the functioning of these tools in a prototypic multiple criteria decision aiding situation; Section 2.4 discusses the relevance of these tools, and motivates their design and use; Section 2.5 identifies venues for improvement of the toolbox, principally geared to address cardinal preferences. Also, not included in this chapter in order to ease reading, Appendix A gathers the proofs of the results at the end of this report.

The tools detailed in Section 2.2, as well as the corresponding proofs in Appendix A, were introduced in [Belahcene et al., 2017a]. All the rest is original material.

2.2 A toolbox for the inference of necessary preferences

The formulation of necessary preference given by Lemma 2.3 is exactly of the form suitable to be processed by Farkas' lemma (Proposition 2.2), with linear forms operating on the vector space of criterion-wise values. Nevertheless, we devote the remainder of the section to the engineering of an encoding of preference as linear forms that is both easier to handle and to interpret.

We proceed as follows: i) we streamline the underlying vector space, in order to obtain a *static* representation of the preference information that does not depend on the alternatives of the query ; ii) we identify trivial cases of impossible necessary preference, and tag them accordingly in order to keep a lean representation of the cases that do matter; and iii) we offer a representation of preference statements in the form of *covectors*, i.e. tuples of coefficients permitting to compute a linear form as an inner product.

The tone of this section is technical, chaining definitions and propositions as it aims at building a repository of tools and providing a convenient reference. Proofs of these results are provided in Appendix A. An illustration and a thorough discussion of these tools can be found in the remainder of this chapter.

2.2.1 Core values and alternatives

We streamline the underlying vector space, in order to obtain a *static* representation of the preference information that does not depend on the alternatives

of the query. To do this, we introduce the notion of *core values*, a tuple of sets of criteria value $\langle \mathbb{P}_i \rangle_{i \in N}$, that encompass the performance values referenced by the preference information according to each point of view, i.e. satisfying the following rule:

$$\mathbb{P}_i \supseteq \bigcup_{(x,y) \in \mathcal{P}} \{x_i\} \cup \bigcup_{(x,y) \in \mathcal{P}} \{y_i\} \tag{2.1}$$

For each point of view, we sort the core values by ascending order of desirability:

$$\langle p_i^1 \prec_i p_i^2 \prec_i \cdots \prec_i p_i^{|\mathcal{P}_i|} \rangle := \mathbb{P}_i \tag{2.2}$$

We call *core alternatives* $\mathbb{P}$ the alternatives built by combining these values.

$$\mathbb{P} := \prod_{i \in N} \mathbb{P}_i \tag{2.3}$$

2.2.2 Unbounded pairs

Beyond monotonicity, we have no information on preferences about values outside of the core. Consequently, there is a class $\mathcal{U}_\mathcal{P}$ of *unbounded* pairs $(x, y) \in \mathbb{X}^2$ that are trivial negative instances for the adjudication of necessary preference problem: when there is a point of view $i \in N$ providing an argument that is both in disfavor of x (because $y_i \succ_i x_i$) and infinitely strong (because $x_i \prec_i \min \mathcal{P}_i$ or $y_i \succ_i \max \mathcal{P}_i$). In such a case, x is clearly *not* necessarily preferred to y.

$$\mathcal{U}_\mathcal{P} := \{(x, y) \in \mathbb{X}^2 \mid \exists i \in N : \left]x_i, y_i\right[\not\subseteq [\min \mathbb{P}_i, \max \mathbb{P}_i]\} \tag{2.4}$$

Theorem 2.4. *:* $\mathcal{U}_\mathcal{P} \cap \mathcal{N}_\mathcal{P} = \varnothing$

Proof. See Appendix A. □

2.2.3 Core intervals and indexes

Consecutive core values define *core intervals* $\langle \left[p_i^k, p_i^{k+1}\right] \rangle_{i \in N, 1 \leq k < |\mathcal{P}_i|}$, indexed by a pair (i, k). We name $\mathbb{I}$ the set of indexes.

$$\mathbb{I} := \bigcup_{i \in N} \{(i, k) \ : \ k \in \mathbb{N} \text{ and } 1 \leq k \leq |\mathcal{P}_i| - 1\} \tag{2.5}$$

2.2.4 Orientation of a core interval w.r.t. a pair of alternatives

Given a pair of alternatives $(x, y) \in \mathbb{X}^2$ and a core index $(i, k) \in \mathbb{I}$ related to the core interval $\left[p_i^k, p_i^{k+1}\right]$, we define a numeric coefficient measuring the

orientation of this particular interval with respect to the comparison between x and y:

$$(x,y)^\star_{(i,k)} := \begin{cases} +1 & \text{, if } [p_i^k, p_i^{k+1}] \subset [y_i, x_i] \\ -1 & \text{, if } [p_i^k, p_i^{k+1}] \cap]x_i, y_i[\neq \varnothing \\ 0 & \text{, else} \end{cases} \tag{2.6}$$

These coefficients partition the intervals $[p_i^k, p_i^{k+1}], (i,k) \in \mathbb{I}$ between pros, cons and neutral arguments of a pair of alternatives (x,y).

2.2.5 Covectors operating on differences of value

Besides their symbolic interpretation, the coefficients measuring the orientation of the core intervals have a numeric role. Given an index $(i,k) \in \mathbb{I}$ and a value function $v_i : \mathbb{P}_i \longrightarrow \mathbb{R}$, we define *elementary differences of value*:

$$\Delta v_{(i,k)} := v_i(p_i^{k+1}) - v_i(p_i^k) \tag{2.7}$$

Therefore, we interpret $(x,y)^\star$ as a covector of $\mathbb{R}^{\mathbb{I}}$, defining a linear form operating on elementary differences of value:

$$(x,y)^\star \cdot \Delta v = \sum_{(i,k)\in\mathbb{I}} (x,y)^\star_{(i,k)} \cdot \Delta v_{(i,k)} \tag{2.8}$$

2.2.6 Representation of dominance through covectors

The canonical dual base is denoted $\mathcal{D}^\star := \langle\ \delta^\star_{(i,k)}\ \rangle_{(i,k)\in\mathbb{I}}$, where the covector $\delta^\star_{(i,k)}$ has all coefficients equal to zero, except for the coefficient associated to the interval indexed by (i,k), which is equal to $+1$, so that $\delta^\star_{(i,k)} \cdot \Delta v = \Delta v_{(i,k)}$

If x dominates y, then all core intervals are oriented nonnegatively with respect to (x,y). Therefore, any dominance statement can be represented by a covector that is a linear combination with nonnegative coefficients of the covectors $\delta^\star_{(i,k)}$:

Lemma 2.5.

The additive values $\langle v_i(p_i^k)\rangle$ *extend* $\mathcal{D} \iff \forall \delta^\star \in \mathcal{D}^\star, \delta^\star \cdot \Delta v \geq 0$ (2.9)

2.2.7 Representation of preference through covectors

For any pair of alternatives in the core $x, y \in \mathbb{P}$, the linear form encoded by the covector $(x,y)^\star$ corresponds exactly to the difference of value between x and y.

Lemma 2.6.

$$\forall x, y \in \mathbb{P},\ V(x) - V(y) = (x, y)^\star \cdot \Delta v, \tag{2.10}$$

where $V = \sum_i v_i$ is the value function representing preference.

Proof. For any core alternative $x \in \mathbb{P}$ and any point of view $i \in N$ we have:

$$v_i(x) = \qquad v_i(p_i^1) + \sum_{k:\ p_i^{k+1} \precsim_i x_i} \left(v_i(p_i^{k+1}) - v_i(p_i^k) \right).$$

Hence, $V(x) = V(p^1) + (x)^\star \times \Delta v$, where $(x)^\star$ is the covector defined over the dual base $\langle \delta^\star_{(i,k)} \rangle_{(i,k)\in\mathbb{I}}$ by:

$$(x)^\star := \sum_{(i,k)\in\mathbb{I}} (x)^\star_{(i,k)} \delta^\star_{(i,k)}, \quad \text{with } (x)^\star_{(i,k)} := \begin{cases} +1, & \text{if } p_i^{k+1} \precsim_i x; \\ 0, & \text{else.} \end{cases} \tag{2.11}$$

Hence, for any pair of alternatives $x, y \in \mathbb{X}$, the differences of values $V(x) - V(y)$ can be expressed as a linear form operating on the core differences Δv:

$$V(x) - V(y) = \left((x)^\star - (y)^\star \right) \times \Delta v$$

When an alternative z belongs to the core, either $z_i \succsim_i p_i^{k+1}$ or $z_i \precsim_i p_i^k$. Equation (2.11) allows to check that, for all $(i,k) \in \mathbb{I}$, $\left((x)^\star - (y)^\star \right)_{(i,k)}$ matches the following value table:

	$x_i \precsim_i p_i^k$	$x_i \succsim_i p_i^{k+1}$
$y_i \precsim_i p_i^k$:	0	+1
$y_i \succsim_i p_i^{k+1}$:	−1	0

Equation (2.6) defining the coefficients of the covector $(x, y)^\star$ allows to check that the coefficients $(x, y)^\star_{(i,k)}$ satisfy the same value table. Therefore, $\left((x)^\star - (y)^\star \right) = (x, y)^\star$.

□

Given some additive values, the preference of x over y is characterized by the nonnegativity of the linear form $(x, y)^\star$ applied to a specific vector of differences of value Δv. Conversely, given some preference information $\mathcal{P} \subset \mathbb{P}^2$, the compatibility of a specific vector of additive values to the preference information can be written:

Lemma 2.7.

$$\textit{The additive values } \langle v_i(p_i^k) \rangle \textit{ extend } \mathcal{P} \iff \forall \pi \in \mathcal{P}, \pi^\star \cdot \Delta v \geq 0 \tag{2.12}$$

2.2.8 Inference of necessary preference through covectors

For pairs outside the class $\mathcal{U}_{\mathcal{P}}$, we give three characterizations of the necessary preference of x over y using covectors.

Theorem 2.8 (characterization of necessary preference using covectors)**.** *Given some preference information $\mathcal{P} \subset \mathbb{P}^2 \subset \mathbb{X}^2$, and a pair of alternatives $(x, y) \in \mathbb{X}^2 \setminus \mathcal{U}_{\mathcal{P}}$, the following propositions are equivalent :*

1. *necessary preference*
$$(x, y) \in \mathcal{N}_{\mathcal{P}}$$

2. *linear feasibility problem*
$$\begin{cases} & (x,y)^\star \cdot \Delta v < 0 \\ \forall \pi \in \mathcal{P}, & \pi^\star \cdot \Delta v \geq 0 \\ \forall (i,k) \in \mathbb{I}, & \delta^\star_{(i,k)} \cdot \Delta v \geq 0 \end{cases} \quad \textit{has no solution } \Delta v \in \mathbb{R}^{\mathbb{I}}$$

3. *combination of statements* $\exists \lambda \in [0, +\infty[^{\mathcal{P}},\ \mu \in [0, +\infty[^{\mathbb{I}}$:
$$(x, y)^\star = \sum_{\pi \in \mathcal{P}} \lambda_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} \mu_{(i,k)} \delta^\star_{(i,k)}$$

4. *integral combination of statements*
$\exists n \in \mathbb{N}^\star,\ \ell \in \mathbb{N}^{\mathcal{P}},\ m \in \mathbb{N}^{\mathbb{I}}$:
$$n\ (x, y)^\star = \sum_{\pi \in \mathcal{P}} \ell_\pi\ \pi^\star + \sum_{(i,k) \in \mathbb{I}} m_{(i,k)}\ \delta^\star_{(i,k)}$$

Proof. : see Appendix A. □

2.3 A working example

While the previous section introduced the mathematical tools underpinning our framework designed to adjudicate necessary preference queries in an interpretable manner, this section illustrates the functioning of these tools, both at the individual level and collectively, by providing an example of the process induced by Theorem 2.8. This workflow is captured and illustrated by figure 2.1. Each box, numbered from one to nine, refers to a specific component of the framework. The remainder of the section details each one of these, building up a continuing example, and providing links to venues for tweaking or augmentation discussed in subsequent sections.

2.3.1 Inputs

The workflow is initiated by the provision of three kinds of input data: the *preference information* describing the attitude of the decision maker, the *core* detailing the values of interest according to each point of view, and the *query* formalizing the question asked to the system.

Component 1 (**Preference information**)**.** As already discussed in Section 1.3.2, the knowledge grounding the preference relation being built is collected in the form of a list of pairwise preference statements, of the type: 'we know that alternative a is at least as good as alternative b'. This knowledge is agnostic to the preference model.

Example. The preference information elicited from the decision maker can be expressed by three preference statements.
$\mathcal{P} := \{\pi_1, \pi_2, \pi_3\}$, with:
π_1 :=((4* , no , 15 min , 180 $) , (2* , yes , 45 min , 50 $))
π_2 :=((4* , no , 45 min , 50 $) , (4* , yes , 15 min , 100 $))
π_3 :=((2* , yes , 15 min , 180 $) , (4* , no , 30 min , 180 $))

> *Going further.* In Section 2.5.2, we discuss the possibility of taking into account a wider spectrum of preference information, e.g. regarding the intensity of preferences.

Component 2 (**Core**)**.** According to each point of view $i \in N$, the core $\mathbb{P}_i$ is the set of attribute values of interest. The core is related to the preference relation by the rule expresses by (2.1).

Example. We opt to define the core on the basis of the values referenced by the preference information:

- from the point of view of comfort:

$$\mathbb{P}_* = \langle p_*^1 := 2^* \prec_* p_*^2 := 4^* \rangle$$

- from the point of view of the presence of a restaurant:

$$\mathbb{P}_r = \langle p_r^1 := \text{no} \prec_r p_r^2 := \text{yes} \rangle$$

- from the point of view of the time spent commuting:

$$\mathbb{P}_t = \langle p_t^1 := 45 \text{ min} \prec_t p_t^2 := 30 \text{ min} \prec_t p_t^3 := 15 \text{ min} \rangle$$

- from the point of view of expenses:

$$\mathbb{P}_\$ = \langle p_\$^1 := 180 \$ \prec_\$ p_\$^2 := 100 \$ \prec_\$ p_\$^3 := 50 \$ \rangle$$

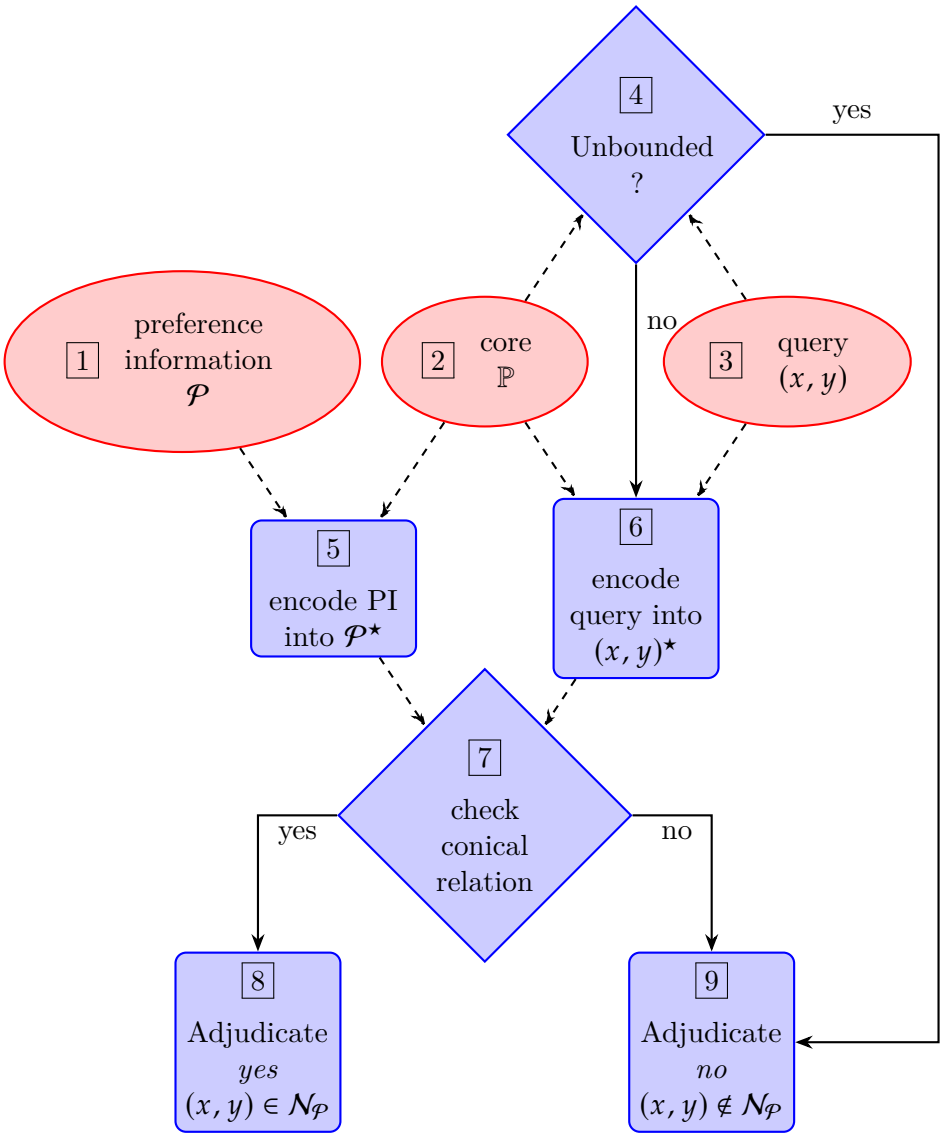

Figure 2.1: Workflow for adjudicating a query under the assumptionof additive preferences

Going further. The relationship between preference information and core can go both ways: obtaining the preference information first, then constructing the core around it, or, conversely, deciding first on the core then collecting corresponding preference statements. These options are further discussed in Section 2.4.1.

Component 3 (Query). The framework is geared towards answering specific

queries, of the form: 'is alternative x preferred to alternative y?'.

Example. We consider the four hotels h_A, h_B, h_C, h_D, whose performance is given by Table 2.1. We suppose the decision maker wants to order them, from the most desirable to the least. Consequently, we are going to consider the twelve queries corresponding to permutations of these hotels.

Going further. In Section 2.4.2 , we show how the framework can be used to efficiently explore the entire preference relation. In Section 2.5.2, we consider answering queries concerning intensities of preference, of the type 'is x preferred to y more strongly than a is preferred to b?'.

2.3.2 The encoder

Taken together, components 4, 5 and 6 form the *encoder*. Their collective function is to represent the *inputs* in a form that is suited to be processed by the inference engine. In our context, inference is linked to the feasibility of a constrained problem, and we represent each piece of preference information by a linear constraint.

Component 4 (**Unboundedness checking**)**.** Unboundedness is a notion defined by Equation 2.4. It qualifies queries that are easy to refute, thanks to Theorem 2.4, because, according to some point of view, the potentially stronger alternative is too weak, or the potentially weaker alternative is too strong.

Example. Among the hotels h_A, h_B, h_C, h_D, two alternatives have an attribute falling outside the range of the core.:

- h_B is further away from the convention center (50 minutes) than the farthest hotel mentioned in the preference information (45 min), and than the other hotels h_A, h_C and h_D. Thus, h_B is not necessarily preferred to these hotels, and the queries $(h_B, h_A), (h_B, h_C)$ and (h_B, h_D) are unbounded.
- h_D is cheaper (40 \$) than the cheapest hotel mentioned in the preference information (50 \$), and than the other hotels h_A, h_B and h_C. Thus, these hotels are not necessarily preferred to h_D, and the queries $(h_A, h_D), (h_B, h_D)$ and (h_C, h_D) are unbounded.

Going further. This notion, its contribution to the interpretability of the framework, and the flexibility it offers, are discussed in Section 2.4.1. This section also discusses how to shut off this bypass entirely.

Component 5 (**Encoding the preference information**)**.** According to Equation 2.12, the linear constraints representing each prefererence statement $(a, b) \in \mathcal{P}$ can be written $(a, b)^\star \cdot \Delta v \geq 0$. The coefficients of the covector $(a, b)^\star$ are given by Equation 2.6. By construction, alternatives mentioned in the preference information are *core alternatives*. For these, Equation 2.6 simply amounts to take into account the core intervals $[p_i^k, p_i^{k+1}]$ between x_i and y_i, positively if $x_i \succ_i y_i$, and negatively if $y_i \succ_i x_i$.

Example. We consider the statement

$$\pi_2 := (\ (4^*, \text{no}, 45\ \text{min}, 50\ \$), (4^*, \text{yes}, 15\ \text{min}, 100\ \$)\ )$$

It expresses the willingness of the decision maker to trade up in cost from 100 $ to 50 $, at the expense of a conjoint downgrade in the presence of a restaurant and in commute time from 15 min to 45 min, everything else —i.e. comfort— being equal. This acceptable trade-off is encoded into a covector given in Table 2.2.

Criteria	Interval	Orientation	Coefficient
Comfort	from 2* to 4*	neutral	0
Restaurant	from no to yes	con	−1
Commute	from 45 min to 30 min	con	−1
	from 30 min to 15 min	con	−1
Cost	from 180 $ to 100 $	neutral	0
	from 100 $ to 50 $:	pro	+1

Table 2.2: Covector encoding the preference satement π_2.

The covectors encoding each preference statements are given in Table 2.3. Those of the dual base, that encode dominance, are given in Table 2.4. Together, they encode the knowledge base.

Criteria	from	to	$\pi_1^\star$	$\pi_2^\star$	$\pi_3^\star$
Comfort	2*	4*	+1	0	−1
Restaurant	no	yes	−1	−1	+1
Commute	45 min	30 min	+1	−1	0
	30 min	15 min	+1	−1	+1
Cost	180 $	100 $	−1	0	0
	100 $	50 $	−1	+1	0

Table 2.3: Covectors encoding the preference information.

Criteria	from	to	$\delta^\star_{*,1}$	$\delta^\star_{r,1}$	$\delta^\star_{t,1}$	$\delta^\star_{t,2}$	$\delta^\star_{\$,1}$	$\delta^\star_{\$,2}$
Comfort	2*	4*	+1	0	0	0	0	0
Restaurant	no	yes	0	+1	0	0	0	0
Commute	45 min	30 min	0	0	+1	0	0	0
	30 min	15 min	0	0	0	+1	0	0
Cost	180 $	100 $	0	0	0	0	+1	0
	100 $	50 $	0	0	0	0	0	+1

Table 2.4: Covectors encoding dominance.

Going further. The way the preference information is encoded, and its consequences is thoroughly discussed in Section 2.4.2. The possibility of representing other types of preference information, e.g. relative to the intensity of preference, is touched upon in Section 2.5.2.

Component 6 (**Encoding the query**)**.** A query $(x, y) \in \mathbb{X}^2$ represents the question of adjudicating whether the alternative x is necessarily preferred to the alternative y, assuming the knowledge contained in the preference information $\mathcal{P}$, and the fact that preferences are additive. Queries that are not unbounded by the core (see Component 4) are encoded in a way that is similar to preference information statements, with covectors $(x, y)^\star$ which coefficients are given by Equation 2.6. As opposed to preference statements, though, preference queries may reference alternatives outside the core. When $(x, y) \notin \mathbb{P}^2$, for some point of view $i \in N$, some attribute x_i, or y_i, or both, falls strictly between the values of $\mathbb{P}_i$, "breaking" some interval $[p_i^k, p_i^{k+1}]$. Because of the cautious nature of the relation $\mathcal{N}_\mathcal{P}$, the orientation of any "broken" interval is rounded down: those that would support the preference of x over y are not taken into account and considered neutral, with coefficient 0, while "broken" intervals that would go against this preference are totally taken into account with coefficient -1. Figure 2.2 illustrates these notions.

Example. The covectors representing the seven queries in $\{h_A, h_B, h_C, h_D\}^2$ that are neither trivial nor unbounded are given in Table 2.5.

Going further. The possibility of efficiently batch querying all the pairs of alternatives from the core is considered in Section 2.4.2. The way the drastically pessimistic attitude implied by the necessary relation manifests when computing covectors of queries mentioning alternatives outside of the core is discussed in Section 2.5.1, which also proposes venues for relaxing it.

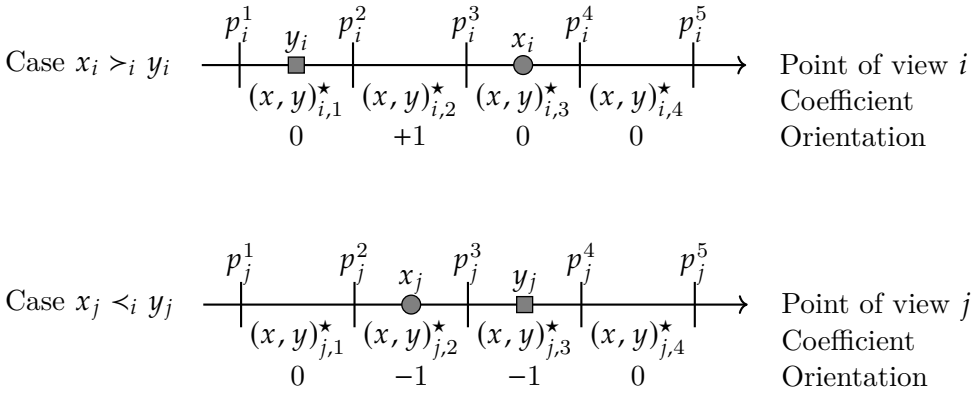

Figure 2.2: Orientation of the core intervals illustrated

Criteria	from	to	$(h_A,h_B)^\star$	$(h_A,h_C)^\star$	$(h_C,h_A)^\star$	$(h_C,h_B)^\star$
Comfort	2*	4*	0	+1	−1	−1
Restaurant	no	yes	−1	−1	+1	0
Commute	45 min	30 min	0	−1	0	+1
	30 min	15 min	0	−1	0	0
Cost	180 $	100 $	0	−1	0	+1
	100 $	50 $	0	−1	+1	+1

Criteria	from	to	$(h_D,h_A)^\star$	$(h_D,h_B)^\star$	$(h_D,h_C)^\star$
Comfort	2*	4*	−1	−1	0
Restaurant	no	yes	0	−1	−1
Commute	45 min	30 min	0	+1	0
	30 min	15 min	0	0	−1
Cost	180 $	100 $	0	+1	0
	100 $	50 $	+1	+1	0

Table 2.5: Covectors encoding the queries.

2.3.3 The inference engine

The inference engine is in charge of adjudication, i.e. determine if a query belongs or not to the necessary preference relation, assuming the preference information and the fact that preferences are additive. As the preference information, previously encoded by Component 5, is treated as an input, it follows that the inference engine embodies the reasoning about preference under the assumption of additivity.

Component 7 (**Computing the adjudication**)**.** The actual computation of the adjudication is done under the auspices of Theorem 2.8, and particularly

the equivalence (1) $\iff$ (3). It amounts to check whether the covector representing the query can be written as a conical combination[c] of the covectors representing the preference information and the covectors of the dual base that represent dominance. Performing such a check is a feasibility problem belonging to the realm of linear optimization.

Example. We have the following relations:

$$(h_C, h_A)^\star = \pi_2^\star + \pi_3^\star + \delta_{r,1}^\star + \delta_{t,1}^\star \tag{2.13}$$

$$(h_C, h_B)^\star = \pi_2^\star + \pi_3^\star + 2\delta_{t,1}^\star + \delta_{\$,1}^\star \tag{2.14}$$

$$(h_D, h_A)^\star = \pi_2^\star + \pi_3^\star + \delta_{t,1}^\star \tag{2.15}$$

Conversely, the covectors $(h_A, h_B)^\star, (h_A, h_C)^\star, (h_D, h_B)^\star$ and $(h_D, h_C)^\star$ can not be expressed as conical combinations of covectors of the knowledge base.

Going further. The fundamental role played by linear programming in the propagation of necessary preference seems to spell the doom of [Spliet and Tervonen, 2014]'s hope of expressing the relation $\mathcal{N}_\mathcal{P}$ in a rule-based framework. Nevertheless, we leave the door open to adjudicating easy *no* instances in a rule-based manner with the bypass offered by the boundedness check (Component 4). Section 2.4.1 offers a discussion of this bypass.

2.3.4 Outputs

As for now, the sole output of the framework is the answer to a single query, either 'yes' or 'no'. In Chapter 3, we will consider the augmentation of the framework with an explicative engine.

Component 8 (**Adjudicating *yes*).** When the Component 7 succeeds at finding a conical relation between the covector representing the query and those representing the preference information and dominance, the alternative x is indeed necessarily preferred to the alternative y: any preference relation based on additive values that extends the preference information deems x at least as good as y.

Example. $(h_C, h_A) \in \mathcal{N}_\mathcal{P}$, $(h_C, h_B) \in \mathcal{N}_\mathcal{P}$ and $(h_D, h_A) \in \mathcal{N}_\mathcal{P}$.

Going further. The explicative engines described in Chapter 3 focus on this case, and make use of the stronger result (1) $\iff$ (4) given by Theorem 2.8: the coefficients of the conical relation can be chosen integral.

[c] A conical combination is a linear combination with nonnegative coefficients.

Component 9 (**Adjudicating *no***)**.** When the query is deemed unbounded by the core by Component 4, or when Component 7 fails at finding a conical relation between the covector representing the query and those representing the preference information and dominance, the alternative x is not necessarily preferred to the alternative y: there is a preference relation based on additive values that extends the preference information and deems y better than x.

Example. $(h_A, h_B) \notin \mathcal{N}_\mathcal{P}$, $(h_A, h_C) \notin \mathcal{N}_\mathcal{P}$, $(h_A, h_D) \notin \mathcal{N}_\mathcal{P}$, $(h_B, h_A) \notin \mathcal{N}_\mathcal{P}$, $(h_B, h_C) \notin \mathcal{N}_\mathcal{P}$, $(h_B, h_D) \notin \mathcal{N}_\mathcal{P}$, $(h_C, h_D) \notin \mathcal{N}_\mathcal{P}$, $(h_D, h_B) \notin \mathcal{N}_\mathcal{P}$ and $(h_D, h_C) \notin \mathcal{N}_\mathcal{P}$.

Going further. In this report, we do not really investigate the particulars of an explanation of the negative cases stemming from a failure of Component 7 to find a conical relation. Section 2.4.1 shortly considers this problem, and also discusses way to increase or decrease the throughput of the bypass offered by Component 4, which can be explained in a rule-based framework.

Epilogue

Example. According to the preference relation $\mathcal{N}_\mathcal{P}$:

- h_A and h_B are incomparable[d];
- h_C is necessarily preferred to h_A and to h_B;
- h_D is necessarily preferred to h_A;
- h_D and h_B are incomparable;
- h_D and h_C are incomparable.

Finally, Figure 2.3 depicts the preference relation $\mathcal{N}_\mathcal{P} \cap \{h_A, h_B, h_C, h_D\}^2$.

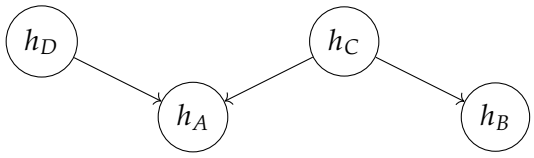

Figure 2.3: Inferred preferences between hotels.

[d]In this model, incomparability has an epistemic flavor. It can be understood as an acknowledgment that more information is required to adjudicate the case.

2.4 What we have done and why it is important

2.4.1 A flexible framework

While the previous section showcased a particular use of the set of tools we put forward in this chapter, this is not the only possible workflow. In this section, we insist on two specific decision points that offer the opportunity to best tailor the framework to the needs of the situation: the acquisition of preference information, and the notion of unbounded pairs.

Core alternatives and preference information. In the example detailed in Section 2.3, we assumed preference information was given at the inset, under the form of pairwise preference statement. We already discussed in Section 1.3.2, the merits of this particular choice for representing the knowledge about preference. We consider here the question of the *initiative* of the protocol governing the acquisition of preference information.

Passive acquisition. In some settings, information is acquired *passively* with respect to the aggregation procedure: observations are made concerning the parameters (supposedly) governing a certain phenomenon, then comes an analyst assuming an additive value model, trying to infer some knowledge about new cases. In this case, corresponding to the example given in Section 2.3, the analyst determines the set of core alternatives after getting the preference information (see Table 2.6) , and the rule 2.1 is read from right to left, as the values mentioned by the preference information constrain the core.

Questioning. In other settings, preference information is learned *actively*: it results from questions asked purposely in order to define the model. This is standard practice in Decision Aiding: the decision maker invites an analyst to help them settle their mind about a 'decision problem', and at some point (see e.g. [Bouyssou et al., 2006]) the analyst proceeds by question the decision maker about their preferences in order to elicit their preference structure. These questions are certainly not randomly chosen, and the values of the attributes mentioned in the queries, i.e. the core $\mathbb{P}$ is crafted by the analyst before the querying, then explored in terms of preference—thus, the axiom 2.1 is read from left to right. For instance, *complete elicitation procedures* aims at collecting preference information sufficient to unequivocally determine the parameters of the aggregation procedure. Of course, as preference information is finite, it imposes to define beforehand the extent to which the model is defined by direct information—in other words, the core—leaving the rest to

A	Collect some preference information
B	**Determine the core**
C	Compute covectors for the preference information

Table 2.6: Workflow for *passive* acquisition of preference information, typical of Machine Learning.

A	**Choose a core**
B	Collect some preference information about core alternatives
C	Compute covectors for the preference information

Table 2.7: Workflow for *active* acquisition of preference information, typical of Decision Aiding.

interpolation (and the choice of an interpolation technique). Realistic methods, that accounts for some amount of inconsistent preference information, such as the Analytical Hierarchic Process [Saaty, 1990] or MACBETH [Bana e Costa and Vansnick, 1994] give explicit guidelines on how to build what we call the core.

Active learning. More recently, *active learning* procedures have been proposed [Braziunas and Boutilier, 2007, Hyafil and Boutilier, 2006, Boutilier et al., 2010, Benabbou et al., 2017] that compute queries on the fly. These queries are presented to the decision maker, who adjudicates them according to their own value, and this adjudication is incorporated into the preference information. The queries are chosen so as to converge quickly towards a sufficiently precise model. These algorithms often work in the version space of the model, as we do, and they might benefit from our efforts towards interpretability and explainability of the recommendations. Their agility requires the core to be computed again with each iteration, but we note that this step does not incur a heavy computational burden[e].

Unbounded pairs bypassing linear optimization. We introduced the notion of unbounded pairs in our framework as a nod towards [Spliet and Tervonen, 2014] efforts to capture necessary preference in a rule-based framework, albeit with very limited results. This introduction is motivated by a concern for interpretability. It offers a shortcut permitting to bypass the linear programming kernel powering the search for certificates of infeasibility. The easy cases are filtered by means of a simple rule-based system, as testing for unboundedness simply asks whether the former alternative of the query is the

[e]Nevertheless, it invalidates some of the benefits of the *streamlined encoding*, discussed in Section 2.4.2.

worst ever seen[f] according to some point of view, or if the latter alternative of the query is the best ever seen according to some point of view. These cases are so clear cut they might deserve a specific treatment. The obvious way of disproving necessary preference—thus explaining the adjudication of rejected queries—is to provide a certificate in the form of additive values amounting to an inversion of preference:

$$(x, y) \notin \mathcal{N}_{\mathcal{P}} \iff \begin{array}{c} \exists V : \mathbb{X} \to \mathbb{R} \text{ additive and extending} \\ \mathcal{P} \text{ and } \mathcal{D} \text{ such that } V(x) < V(y) \end{array} \tag{2.16}$$

Choosing and displaying such an adversarial additive value V is not trivial, probably involving mathematical programming for the artificial analyst on the emitting end and some cognitive burden for the decision maker on the receiving end. Moreover, the less obvious pitfall hides in the anecdotal value of such an adversarial certificate. It surely proves the inadequacy of x compared to y, but it does not tell us anything obviously useful when comparing z to y or x to t, even less z to t. This is a limitation inherent to transductive approaches, inference from particular to particular[g]. Nevertheless, the notion of unbounded pairs circumvents this failure to generalize, by providing a universal and synthetic rule for some cases.

Unbounded pairs are not a vital component of the framework, and can easily and harmlessly be switched off. This can be done by augmenting the sets of core values: from above with an ideal element comparing favorably to any other, and from below with an anti-ideal element comparing unfavorably to any other. In such a setting, there are no unbounded pairs, at the extra cost of having two extra dimensions for each point of view. Conversely, the notion of unbounded pairs could easily be expanded, e.g. by considering separate core scales for the first and second arguments of the preference statements, in order to extend the domain of rule-based exceptions to the linear feasibility framework.

2.4.2 A streamlined representation

A static representation of knowledge. The only alternatives that matter in the definition of necessary preference under the assumption of additive values are: i) those referenced by the preference information; and ii) those that are currently queried. Therefore, the linear program proposed in [Greco et al., 2008], recalled in Proposition 2.1, references these two types of values, and works perfectly fine. From the perspective of the representation of knowledge

[f] In the sense of the alternatives of the query as well as those of the preference information, but not those encountered e.g. during previous queries.

[g] See Section 2.4.2 for a more detailed discussion of inductive and transductive approaches

and reasoning, though, this situation was not quite satisfying. Preference information makes up for the *knowledge base*, the known facts about the world[h], while the adjudication of queries is the *inference engine*, reasoning about those facts and using some theorem to deduce new facts. Former ways of adjudicating queries did not establish as clear a separation between knowledge base and preference engine as the one proposed here. Moreover, the toolbox presented in section 2.2 proposes a way of representing facts (preference statements in $\mathcal{P}$) and queries alike in a static manner, i.e. independently from the queries themselves. In a sense, knowledge is *compiled* beforehand into covectors, and queries are adjudicated at runtime by the inference engine. This feature identifies the inference engine as a transductive-inductive hybrid:

- *transductive*, defined by [Gammerman et al., 1998b, Pirlot et al., 2016, Russell, 1912] as inference from particular to particular, because each individual fact—a single preference statements—is encoded into its own covector, making it possible to trace its particular influence on each decision[i];
- *inductive*, because the inference engine embodies the adjudication rule, tying particular runs—adjudication of queries—together into a general model. This would not have been the case if the encoding of facts had been dynamic—each run would take place in a different landscape, shutting off any possibility for the observer to identify patterns.

Few coefficients and a clear semantic. The encoding we propose uses one coefficient, with value -1, 0 or 1, for each core interval, that is the number of core values minus one for each point of view. This number is slightly less than those required in Proposition 2.1 or Lemma 2.3 in the worst case, which is the number of core values plus two for each point of view. This improvement comes from the outer treatment of unbounded pairs. We believe its impact to be modest on the computational resources required to adjudicate a single query, as linear optimization is polynomial time. Where the novel encoding shines, though, is the way it enables the users (decision maker and analyst alike) to get a global view of the necessary preference relation. Its influence is twofold:

- The computation of covectors (with equation 2.6) comes with a very simple interpolation rule, resorting to logic rather than arithmetic, permitting to limit the scope of exploring the binary relation $\mathcal{N}_\mathcal{P}$ from the

[h] Besides preference information, there is another important fact: *preferences are additive.*

[i] This would not have been the case if the facts had been digested into parameters of a model, such as the (in)famous *weights denoting the importance of criteria* often discussed in Multiple Criteria Decision Aiding.

whole set of pairs of alternatives $\mathbb{X}^2$ to the finite, but combinatorial, set of pairs of core alternatives $\mathbb{P}^2$;

- Rather than reasoning in terms of pairs of alternatives (i.e. $\mathbb{P}^2$, it is sufficient to run through the set of their covectors, which is a subset of $\{-1, 0, 1\}^{\mathbb{I}}$).

The latter point needs to be examined closely, as the computational gain is far from obvious. The mapping from pairs of core alternatives to covectors is neither injective nor surjective.

- There are pairs of core alternatives that are represented by the same covector: the case where, according to some point of view, every coefficient is null, arises if and only if the alternatives of the pair share a common value of the attribute corresponding to this point of view, regardless of the particular value. Thus, the encoding takes advantage of the *preferential independance* property of the additive value model, and represent all the *ceteris paribus* pairs in the same manner.

- Conversely, there are tuples in $\{-1, 0, 1\}^{\mathbb{I}}$ that do not represent any pair of alternatives. It follows from equation 2.6 that: i) each point of view $i \in N$ is either in favor of the pair, with coefficients 0 or +1; completely neutral, with all coefficients null; or in disfavor of the pair, with coefficients 0 or −1; and ii) for each point of view $i \in N$ the set of nonzero coefficients is an interval. It is also easy to find, for any covector satisfying (i) and (ii), a pair of alternatives that is encoded by this covector. Therefore, it is possible to count the covectors encoding pairs of alternatives.

  Lemma 2.9.

$$\left| \bigcup_{(x,y)\in\mathbb{P}^2} \{(x, y)^\star\} \right| = \prod_{i\in N} (|\mathbb{P}_i| \times (|\mathbb{P}_i| - 1) + 1) \tag{2.17}$$

  Proof. From the discussion above, there are as many covectors encoding pairs of alternatives as there are products of intervals with a given orientation. For a given point of view $i \in N$, there are $\binom{|\mathbb{P}_i|}{2}$ intervals pro, $\binom{|\mathbb{P}_i|}{2}$ intervals con, and one neutral interval. □

- Moreover, it is noteworthy that some of the meaningful covectors encode dominance relationships. They are both easy to characterize— they are the covectors with either no pros (in $\{-1, 0\}^{\mathbb{I}}$) or no cons (in $\{0, +1\}^{\mathbb{I}}$), with the null covector as sole overlap—and to adjudicate.

Example 2.2. Therefore, the necessary preference relation given the core chosen for the example given in Section 2.3 involves $|\mathbb{P}_*| \times |\mathbb{P}_r| \times |\mathbb{P}_t| \times |\mathbb{P}_\$| =$

$2\times2\times3\times3 = 36$ core alternatives. Instead of adjudicating $36\times36 = 1296$ pairs of alternatives, it is sufficient to adjudicate $(|\mathbb{P}_*|\times(|\mathbb{P}_*|-1)+1)\times(|\mathbb{P}_r|\times(|\mathbb{P}_r|-1)+1)\times(|\mathbb{P}_t|\times(|\mathbb{P}_t|-1)+1)\times(|\mathbb{P}_\$|\times(|\mathbb{P}_\$|-1)+1) = 3\times3\times7\times7 = 441$ covectors. Among these, $(1+1)\times(1+1)\times(3+1)\times(3+1) = 64$ correspond to covectors without any cons, encoding pairs $(x, y) \in \mathcal{D}$ and are trivial *yes* queries, and the same number correspond to covectors without any pros encoding pairs (x, y) where $y\ \mathcal{D} x$, and are therefore trivial *no* queries except in the only overlapping case where $x \equiv y$ and $(x, y)^\star$ is the null covector. Thus, there remains $441 - 2\times64 + 1 = 314$ nontrivial covectors to adjudicate. Of these, 53 are *yes* instances. Finally, among the 1296 pairs of alternative, 288 are ordered by dominance, and 250 more are ordered by the necessary preference relation: not too bad, considered there are only three preference statements! Figure 2.4 depicts the Hasse diagram, with 36 nodes and 57 edges, of the relation $\mathcal{N}_\mathcal{P}$, i.e. the graph of $\mathcal{N}_\mathcal{P}$ where arcs that can be deduced from transitivity are omitted.

2.5 Perspectives and venues for improvement

This section is devoted to the description of two venues for increasing the expressiveness of the framework presented in this chapter, thus extending the scope of preferences it is able to represent. The first one, detailed in Section 2.5.1, aims at equipping the framework with the ability to represent assumptions made about the shape of the marginal value functions, frequently encountered in the multiple criteria decision aiding literature. The second, detailed in Section 2.5.2, considers the issue of representing the intensity of preference and regret inside the framework.

2.5.1 Relaxing orientations to account for interpolation

While the multiple criteria decision aiding community—researchers and practitioners alike—considers the additive value model as a flagship, it is seldom used *as is* in practical applications. Additional assumptions are often made concerning the shape of the marginal values, i.e. the functions $\langle v_i : \mathbb{X} \to \mathbb{R}\rangle_{i\in N}$. In the framework presented this far, these functions play a purely *ordinal* role: the order of the numeric values $\langle v_i(x)\rangle_{x\in\mathbb{X}}$ reflects the preference $\succsim_i$ over the alternatives of $\mathbb{X}$. The current framework makes no assumption about the shape of these functions, besides the fact they are monotonically nondecreasing. When there is no assumption made about the shape of the marginal values, necessary preference commands to interpolate values in the most demanding way, expressed by the *pessimistic rounding rule*: considering the core intervals as *assets*, a potentially pro interval is actually counting as supportive only if it

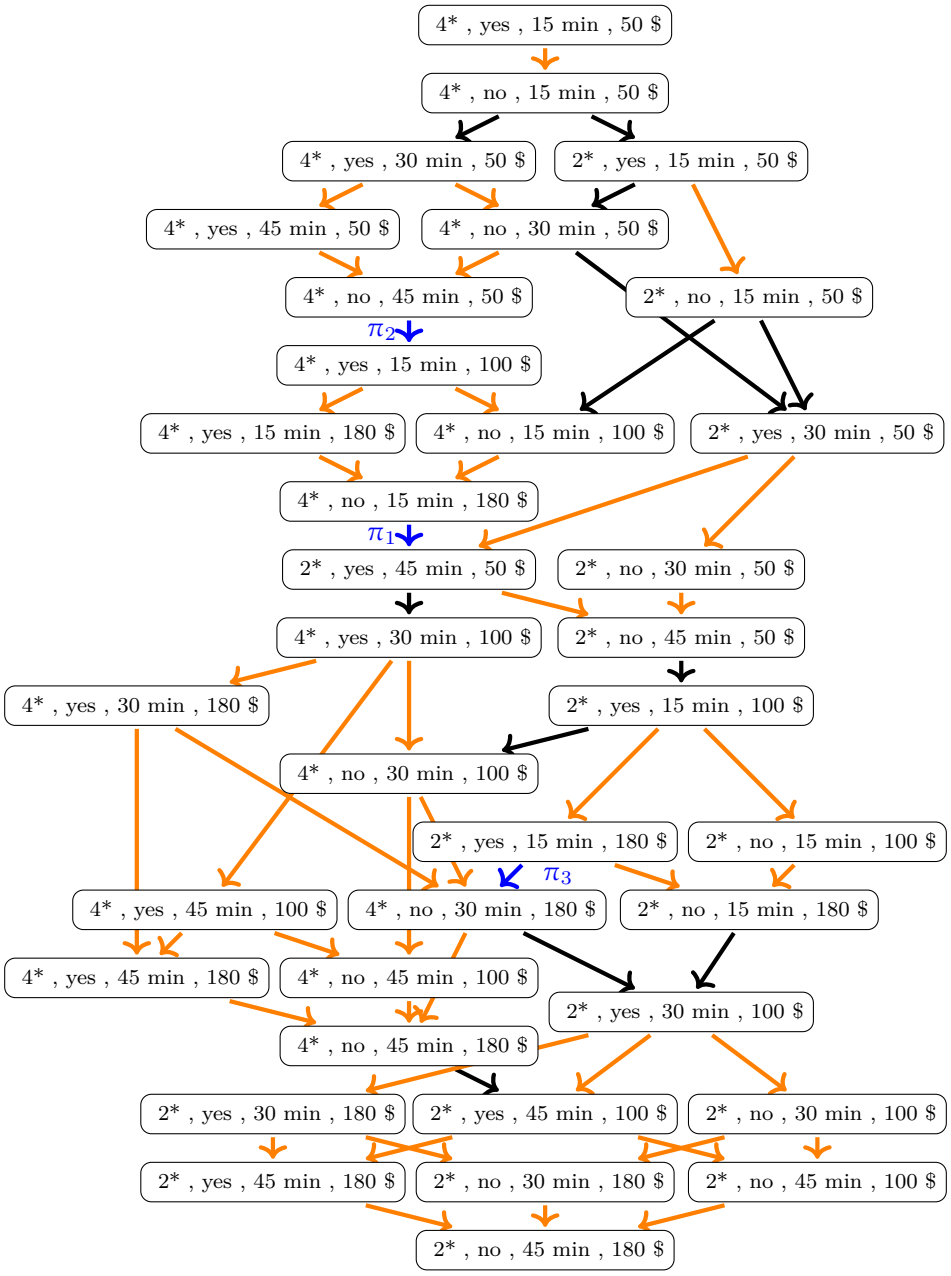

Figure 2.4: Hasse diagram of the necessary preference relation $\mathcal{N}_{\mathcal{P}}$. Preference information is in blue, dominance is in orange, and inferred relations are in black.

is entirely covered by the interval ranging from y_i to x_i, with partial cover not counting at all; conversely, a potentially con interval counts as full opposition as soon as it intersects the interval ranging from x_i to y_i.

Example 2.3. Consider a situation where, from some point of view $i \in N$, the core is made of four attribute values $p_i^1 \prec_i p_i^2 \prec_i p_i^3 \prec_i p_i^4$, and alternatives x and y such that $p_i^1 \prec_i x \prec_i p_i^2$ and $p_i^3 \prec_i y \prec_i p_i^4$. Figure 2.5 depicts two value functions:

- one corresponding to a case extremely in disfavor of x compared to y, with $v_i(x) = v_i(p_i^1)$ and $v_i(y) = v_i(p_i^4)$. Thus, $v_i(x) - v_i(y) = (-1) \times (v_i(p_i^2) - v_i(p_i^1)) + (-1) \times (v_i(p_i^3) - v_i(p_i^2)) + (-1) \times (v_i(p_i^4) - v_i(p_i^3))$, encoded into the covector $(-1, -1, -1)$
- one corresponding to a case extremely in disfavor of y compared to x, with $v_i(x) = v_i(p_i^2)$ and $v_i(y) = v_i(p_i^3)$. Thus, $v_i(x) - v_i(y) = (0) \times (v_i(p_i^2) - v_i(p_i^1)) + (+1) \times (v_i(p_i^3) - v_i(p_i^2)) + (0) \times (v_i(p_i^4) - v_i(p_i^3))$, encoded into the covector $(0, +1, 0)$

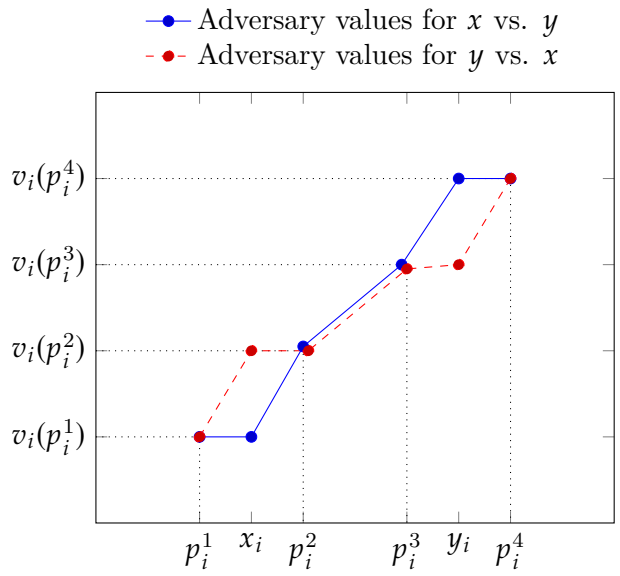

Figure 2.5: Most disfavorable values without assumptions w.r.t. interpolation

Nevertheless, many practical frameworks dedicated to the elicitation of preferences make additional assumptions on the marginal values:

- Weighted sums, certainly the most popular way of aggregating values without bothering with re-encoding them, correspond to the case where marginal values are linear.

- UTA [Jacquet-Lagrèze and Siskos, 1982] or MACBETH [Bana e Costa and Vansnick, 1994], marginal values are assumed to be piecewise linear.
- More recently, [Benabbou et al., 2016, Sobrie et al., 2018] assume piecewise polynomials.
- Decision theoretic considerations could furthermore constrain the values, e.g. forcing concavity in the case of diminishing returns.

All of these assumptions, and more, can be integrated in our framework by modifying solely the *encoder*, but not the component dedicated to adjudication, nor the explicative components introduced in Chapter 3. As a proof of concept, we give two working examples showing how the coefficients of the covector representing a query can be adjusted in order to account for additional assumptions.

Example 2.4. Suppose the values are piecewise linear, with breaking points at the core values, in the manner of weighted sums (with a single core interval), UTA [Jacquet-Lagrèze and Siskos, 1982] or MACBETH [Bana e Costa and Vansnick, 1994]. Consider the values $p_i^1 \prec_i x \prec_i p_i^2 \prec_i p_i^3 \prec_i y \prec_i p_i^4$ as depicted in example 2.3. Now, the contribution of the core interval $[p_i^1, p_i^2]$ to the value of x is known, equal to $\frac{x-p_i^1}{p_i^2-p_i^1} \times (v_i(p_i^2) - v_i(p_i^1))$. In the same vein, $v_i(y) = v_i(p_i^3) + \frac{y-p_i^3}{p_i^4-p_i^3} \times (v_i(p_i^4) - v_i(p_i^3))$ Hence, the queries (x, y) and (y, x) can be represented, according to the point of view $i \in N$, by the respective covectors

$$(x, y)_i^{\star \text{ piecewise linear}} := (-1 + \frac{x - p_i^1}{p_i^2 - p_i^1}, -1, -1 + \frac{p_i^3 - y}{p_i^4 - p_i^3})$$
$$(y, x)_i^{\star \text{ piecewise linear}} := (0 + \frac{p_i^2 - x}{p_i^2 - p_i^1}, +1, 0 + \frac{y - p_i^3}{p_i^4 - p_i^3})$$

Figure 2.6 illustrates this situation.

Example 2.5. Suppose the values are constrained in an envelope such that, given $p_i^1 \prec_i x \prec_i p_i^2 \prec_i p_i^3 \prec_i y \prec_i p_i^4$, $\underline{\kappa_i(x)} \times \Delta v_i^1 \leq v_i(x) - v_i(p_i^1) \leq \overline{\kappa_i(x)} \times \Delta v_i^1$ and $\underline{\kappa_i(y)} \times \Delta v_i^3 \leq v_i(y) - v_i(p_i^3) \leq \overline{\kappa_i(y)} \times \Delta v_i^3$, where the coefficients $\underline{\kappa_i(x)}, \overline{\kappa_i(x)}, \underline{\kappa_i(y)}, \overline{\kappa_i(y)}$ can be computed beforehand— and accounted for separately. Then, the necessary preference of x over y or of y over x can be computed with Theorem 2.8, with covectors corresponding to the most unfavorable case:

$$(x, y)_i^{\star \text{ envelope}} := (-1 + \underline{\kappa_i(x)}, -1, -\overline{\kappa_i(y)})$$
$$(y, x)_i^{\star \text{ envelope}} := (1 - \overline{\kappa_i(x)}, +1, \underline{\kappa_i(y)})$$

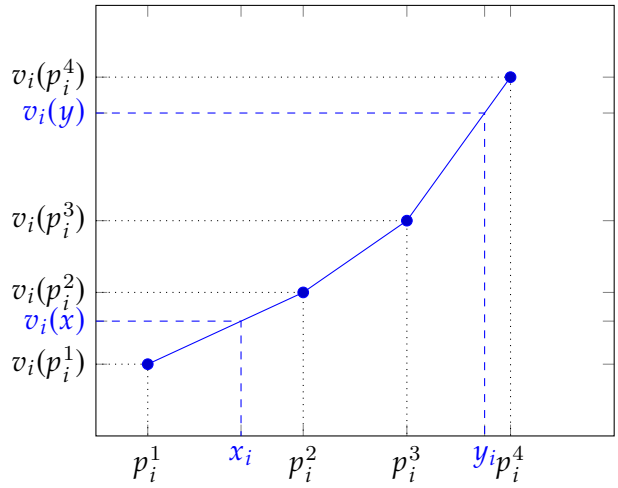

Figure 2.6: Values with linear interpolation.

Figure 2.7 illustrates this situation.

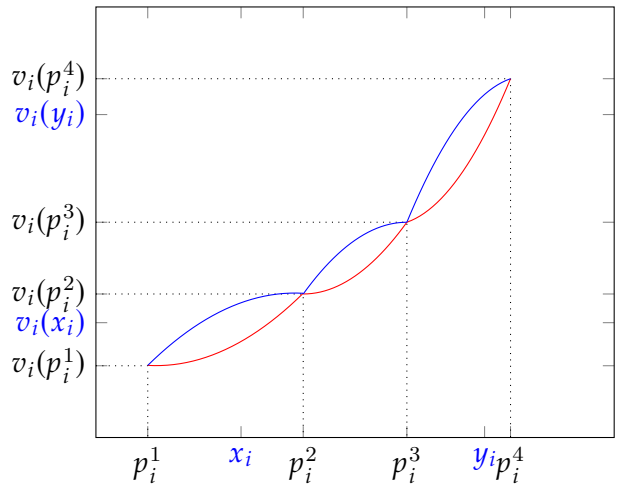

Figure 2.7: Envelopes of values.

The case where both x and y fall inside the same core interval needs to be addressed with extra care, and is not discussed in this report. Moreover, as long as the values of the coefficients $\underline{\kappa_i(x)}, \overline{\kappa_i(x)}, \underline{\kappa_i(y)}, \overline{\kappa_i(y)}$ are rational, the characterization of necessary preference via a conical relation with integral coefficients (Theorem 2.8, (1) $\iff$ (4)), that serves as a basis for the explanation tools propose in Chapter 3, holds.

2.5.2 Encoding the intensity of preferences

In [Figueira et al., 2009], the notion of necessary preference assuming additive values is extended to account for intensity of preferences. Formally, the intensity of preference can be introduced as a binary relation between pairs of alternatives (that can therefore be seen as a quaternary relation between alternatives), with the following semantic: given two pairs of alternatives $(a,b),(c,d)$ such that a is preferred to b and c is preferred to d, the pair (a,b) is in relation to the pair (c,d) if a is preferred to b more intensely than c is preferred to d. This relation enriches the language of preference in a manner that seems easy to grasp in a decision aiding context, and methods such as MACBETH [Bana e Costa and Vansnick, 1994] make heavy use of this kind of statements as a source of preference information. They are also easy to model in a value-based framework[j], with the following rule: the preference (a,b) is more intense than the preference (c,d) if, and only if, the difference of value $V(a)-V(b)$ is greater than the difference of value $V(c)-V(d)$.

This rule is easy to represent in the present framework, as $(V(a))-V(b))-(V(c)-V(d))$ is clearly a linear form of the elementary differences of preference Δv. Hence, an intensity of preference statement can be represented by a covector, and be incorporated into the knowledge base if it is an acknowledge fact, or queried and adjudicated in order to check if this intensity can be inferred from the knowledge base. As the coefficients of these covectors are clearly integral, Theorem 2.8 should require very little modification to account for intensities of preference.

At a formal level, this interpretation of differences of differences of value as intensities rule is motivated by the fact that, in the context of additive conjoint measurement, the value scales constructed by the measurement process are *interval scales*, i.e. defined up to a positive affine transformation. These transformations do not affect the orientation of the differences of differences of value, which are therefore invariant with respect to the arbitrariness of the scale (its position and magnitude). In the specific context considered here, though, the extent to which the intensity of preference can reliably be assessed is not clear. This context is generally labeled as *ordinal*, as the only informational primitives it relies on are orientations of preference, not magnitude. It is therefore generally considered unsuitable to the derivation of *cardinal* information tied to the magnitude of the values. To this negative consideration, it can be objected that: i) if the orientation of comparisons between intensities of preference is based on the *sign* of differences of differences

[j]This type of reasoning can also be extended to non-numeric representations, following the principle of *analogical proportion*—see e.g. [Prade and Richard, 2018, Bounhas et al., 2018].

of value, it can still be considered as ordinal, as it falls clearly under the umbrella of the Farkas-based framework hereby described; and ii) the extent to which this type of information can be derived from the current informational state of an incompletely elicited model is exactly what the computation of the necessary preference relation is about. Consequently, we expect the tools introduced here to help further the understanding of this particular issue.

A second step can be considered towards cardinality, in order to augment the expressiveness of the framework. It is well-known that necessary preference is a demanding notion, and may result in a preference structure that is very sparse besides dominance. A natural way to extend the notion of necessary preference is to interpret differences of values in terms of pairwise *regret* [Savage, 1951]—the loss of value incurred by choosing an alternative instead of another. In the context of incomplete information, when an alternative is necessarily preferred to another, the regret of choosing it is negative, whatever the 'state of the nature'— here, the *ground truth*, i.e. the value model describing the preference structure. Maybe, when things are not so clear cut, because preference information is lacking, no alternative is necessarily preferred to the other. In order to qualify this case more precisely, it might be interesting to have a look at the pairwise maximal regret. This cardinal information has been fruitfully used in several *active learning* frameworks for the elicitation of various models [Braziunas and Boutilier, 2007, Boutilier et al., 2006] or, more recently, [Benabbou et al., 2017]. We remark that, in order to give an interpretation to regret, one needs to normalize the value scales. This remark is key to incorporating regret into our framework: any 'pairwise regret statement' can be formulated as a statement concerning intensity of preferences. For instance, stating that the maximum pairwise regret incurred by choosing x instead of y is five percent can be written as:

$$V(y) - V(x) \leq \frac{5}{100}(V(\top) - V(\bot))$$

or, equivalently:

$$100(V(y) - V(x)) \leq 5(V(\top) - V(\bot)),$$

where $\top$ and $\bot$ are, respectively, the ideal and anti-ideal alternatives, with maximal (resp. minimal) value. As long as the maximum pairwise regret is a rational number, any pairwise regret statement can be expressed as linear form operating on elementary differences of value, and encoded into a covector with integral coefficients.

3

Explanations

Introduction

This chapter is devoted to the presentation of elements of explanation supporting additive preferences. This explanatory step comes after the computation step, where necessary preference statements are adjudicated, detailed in Chapter 2. The explanation engines developed in this chapter abundantly leverage the Farkas certificate with integer coefficients (see theorem 2.8, fourth formulation) to provide argumentative support to inferred preference statements. We propose two different approaches to the question of explaining an inferred preference.

- The first one, described in Section 3.1, is concerned with putting in relation specific statements of the preference information, seen as *premises*, to the inferred statement, seen as a *conclusion*. It operates by interpreting a Farkas certificate as a pointer towards relevant preference statements, where the pros are known to be stronger than the cons, then carving out the desired conclusion by means of a *cancellative* property.

- The second one, described in Section 3.2, is based on *sequences of preference swaps*, inspired by the *even-swaps* method for the active learning of an additive model, proposed in [Hammond et al., 1998]. The preference statement that requires an explanation is tentatively broken down into a chain of transitive statements, supposedly easy to accept (or refute) because they are restricted to the expression of trade-offs between two points of view.

3.1 Explaining via cancellation

This section is dedicated to the presentation of an explanation technique. We provide argumentation schemes [Walton and Reed, 2002] for explaining pairwise preference statements entailed from a robust additive value model. This model is a common way of aggregating preferences derived from incomplete preference information stemming from multiple points of view. We ground the arguments schemes on a cancellation principle, and prove that these explanations are necessary and sufficient conditions to support necessary preferences under the assumption of the additive value model. We also prove that, while the inference is polynomial time, finding a cancellative explanation is NP-complete.

3.1.1 Cancellative properties of the additive value model

Besides *transitivity*, *completeness* and extending dominance, the additive value model has many specific properties. The most salient, though, are certainly the *cancellative* ones. Their definitions can be found e.g. in [Krantz et al., 1971, Wakker, 1989]. They are easier to describe using a syntactic facility: for any nontrivial subset of points of view $A \subset N$ and any two alternatives $a, b \in \mathbb{X}$, we denote $a_{-A}b_A$ the (fictitious) alternative which is equivalent to a according to each point of view not in A, and equivalent to b according to the points of view in A. We also define the sets of shared and differing attributes between two alternatives:

$$\forall x, y \in \mathbb{X}, \quad N^{=}_{(x,y)} := \{i \in N : x \sim_i y\} \tag{3.1}$$

$$N^{\neq}_{(x,y)} := \{i \in N : x \nsim_i y\} \tag{3.2}$$

First-order cancellation. This property can be described as the independence of the aggregated preference to indifferent points of view. Formally, $\forall A \subset N, A \neq \emptyset, \forall x, y, z, z' \in X$:

$$x_{-A}z_A \succsim y_{-A}z_A \iff x_{-A}z'_A \succsim y_{-A}z'_A \tag{3.3}$$

This property formalizes the possibility to reason *ceteris paribus*—everything else being equal. When a preference relation satisfies first-order cancellation, the specific levels of the attributes in $A \equiv N^{=}_{(x,y)}$ are irrelevant to the adjudication of the preference between x and y, and thus does not need to be mentioned. This certainly opens up opportunities w.r.t. explanation.

Higher-order cancellations These properties involve canceling out terms across multiple pairwise preference statements. They have been extensively

studied, in relation with the axiomatization of the additive model, in e.g. [Scott, 1964, Krantz et al., 1971, Fishburn, 1997]. They are easier to express with a permutation, e.g., for any given integer $m \geq 1$, the *mth-order cancellation axiom:*

> Consider $m+1$ alternatives $x^{(0)}, \ldots, x^{(m)}$ in $\mathbb{X}$. Let $y^{(0)}, \ldots, y^{(m)}$ be $m+1$ alternatives in $\mathbb{X}$ such that, for every point of view $i \in N$, $(y_i^{(0)}, \ldots, y_i^{(m)})$ is a permutation of $(x_i^{(0)}, \ldots, x_i^{(m)})$. Then,$[x^{(k)} \succsim y^{(k)}, \forall k \in \{1, \ldots, m\}] \Rightarrow y^{(0)} \succsim x^{(0)}$ (3.4)

The following—obvious—theorem connects these properties to the additive value models and, more importantly, to their robust counterpart.

Theorem 3.1. *Any additive value model satisfies the cancellation axiom at any order. Given some preference information $\mathcal{P} \subset \mathbb{X}^2$, the necessary preference relation assuming additive values $\mathcal{N}_{\mathcal{P}}$ satisfies the cancellation axiom at any order.*

3.1.2 Syntactic cancellation

Maybe the reasoning illustrated by Example 2.1 is exactly the prototype of the explanations we are looking for. We formalize it under the form of an *argument scheme* [Walton, 1996], an operator tying *premises* satisfying some conditions, to a *conclusion.*

Definition 3.1 (syntactic cancellative explanation scheme)**.** *Given two positive integers $m \geq n$, and a pair of alternatives $(x, y) \in \mathbb{X} \times \mathbb{X}$, we say the tuple $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathbb{X} \times \mathbb{X})^m$ is a* syntactic cancellative explanation *of length m with n repetitions of the pair (x, y) if, for each point of view $i \in N$, $(\underbrace{y_i, \ldots, y_i}_{n \text{ repetitions}}, a_i^{(1)}, \ldots, a_i^{(m)})$ is a permutation of $(\underbrace{x_i, \ldots, x_i}_{n \text{ repetitions}}, b_i^{(1)}, \ldots, b_i^{(m)})$.*

Verification of an explanation. Checking if a given tuple of pairs of alternatives is an explanation of a given pair of alternatives with a given number of repetitions can be performed in $O(|N| \cdot m \ln m)$ by Algorithm 3.1. At the heart of this algorithm, the unitary check performed by line 7, made at the level of the attribute values, is indeed syntactic.

Certificate of necessary preference. The purpose of these argument schemes is to provide a certificate for necessary preference.

Algorithm 3.1: Checking a syntactic cancellative explanation

Input:
- a pair of alternatives $(x, y) \in \mathbb{X}^2$;
- a tuple of pairs of alternatives $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathbb{X} \times \mathbb{X})^m$;
- a positive integer n.

Result: *True*, if the tuple $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle$ is a syntactic cancellative explanation with n repetitions of the pair (x, y); *False* else.

```
foreach point of view i ∈ N: do
    create a list L_i of length m + n containing the values a_i^(1) ... a_i^(m)
      as well as n copies of y_i;
    sort L_i in ascending order according to ≿_i;
    create a list R_i of length m + n containing the values b_i^(1) ... b_i^(m)
      as well as n copies of x_i;
    sort R_i in ascending order according to ≿_i;
    foreach index k ∈ [1..m + n] do
        if L_i[k] ≺_i R_i[k] then
            return False;
return True.
```

Theorem 3.2. *Given some preference information $\mathcal{P} \subset \mathbb{X}^2$ and two alternatives $x, y \in X$, $(x, y) \in \mathcal{N}_{\mathcal{P}}$ if, and only if, there are two positive integers m and n and a tuple of pairs of alternatives $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathcal{P} \cup \mathcal{D})^m$ forming a cancellative explanation of length m with n repetitions of the pair (x, y).*

This theorem is a more specific version of the result given, with a similar[a] sketch of proof in [Wakker, 1989] (II.3 on p33). For our purpose, the coefficient n, ignored by Wakker, can not be disregarded as unimportant.

Proof. We propose a direct proof, similar to the one sketched in [Wakker, 1989] (II.3 on p33), in Appendix A. Yet, the other results obtained in this chapter make this proof redundant, as illustrated by Figure 3.1.

- If there is a cancellative explanation with n repetitions with pairs in the preference information and dominance relation, then, for any ad-

[a] Farkas' lemma and the hyperplane separation theorem are, essentially, the same property with different names.

ditive value function $V = \sum_{i \in N} v_i \ : \mathbb{X} \to \mathbb{R}$ compatible with dominance $\mathcal{D}$ and the preference information $\mathcal{P}$, summation over the tuples $(\underbrace{y_i, \ldots, y_i}_{n \text{ repetitions}}, a_i^{(1)}, \ldots, a_i^{(m)})$ on the one hand and $(\underbrace{x_i, \ldots, x_i}_{n \text{ repetitions}}, b_i^{(1)}, \ldots, b_i^{(m)})$ on the other hand yields, for all points of view $i \in N$:

$$n \cdot v_i(y) + \sum_{k=1}^{m} v_i(a^{(k)}) = n \cdot v_i(x) + \sum_{k=1}^{m} v_i(b^{(k)})$$

Then, summation over points of view yields

$$n \cdot V(y) + \sum_{k=1}^{m} V(a^{(k)}) = n \cdot V(x) + \sum_{k=1}^{m} V(b^{(k)})$$

Therefore

$$V(x) - V(y) = \frac{1}{n} \sum_{k=1}^{m} \left(V(a^{(k)}) - V(b^k) \right)$$

Each term of the sum is a difference of values known to be nonnegative and the coefficient n is positive, thus the difference $V(x) - V(y)$ is nonnegative and x is necessarily preferred to y.

- Reciprocally, if $(x, y) \in \mathcal{N}_{\mathcal{P}}$, Farkas' lemma (Proposition 2.2) applied to the canonical representation of necessary preference (Lemma 2.3) yields a certificate of the form:

$$\sum_{i \in N} v_i(x) - \sum_{i \in N} v_i(y) = \sum_{(a,b) \in \mathcal{P} \cup \mathcal{D}} \lambda_{(a,b)} \cdot \left(\sum_{i \in N} v_i(a) - \sum_{i \in N} v_i(b) \right) \tag{3.5}$$

satisfied by any additive value function compatible with dominance $\mathcal{D}$ and the preference information $\mathcal{P}$. As the coefficients of the linear forms are integers (indeed, there are in $\{-1, 0, +1\}$), the nonnegative coefficients $\lambda_{(a,b)}$ can be chosen rational. Multiplying both sides of (3.5) by a common multiple of their denominators yield a conical combinations with integral coefficients:

$$n \cdot \left(\sum_{i \in N} v_i(x) - \sum_{i \in N} v_i(y) \right) = \sum_{(a,b) \in \mathcal{P} \cup \mathcal{D}} \ell_{(a,b)} \cdot \left(\sum_{i \in N} v_i(a) - \sum_{i \in N} v_i(b) \right) \tag{3.6}$$

Equation (3.6) can be rewritten as:

$$\sum_{i \in N} \left(n \cdot v_i(x) + \sum_{(a,b)} \ell_{(a,b)} \cdot v_i(b) \right) = \sum_{i \in N} \left(n \cdot v_i(y) + \sum_{(a,b)} \ell_{(a,b)} \cdot v_i(a) \right) \tag{3.7}$$

This is a vector equality, between vectors indexed by a pair $(i,z) \in \bigcup_{i\in N} \bigcup_{z\in\widehat{\mathbb{P}}_i} \{(i,z)\}$, with the *augmented core*[b] defined by:

$$\widehat{\mathbb{P}}_i := \bigcup_{(a,b)\in\mathcal{P}} \{a_i\} \cup \bigcup_{(a,b)\in\mathcal{P}} \{b_i\} \cup \{x_i\} \cup \{y_i\}; \tag{3.8}$$

Consider the tuple E of length $m := \sum_{(a,b)\in\mathcal{P}\cup\mathcal{D}} \ell_{(a,b)}$ formed by concatenating all the pairs (a,b) appearing in this sum with a nonzero $\ell_{(a,b)}$ coefficient, each pair (a,b) being repeated $\ell_{(a,b)}$ times. For any point of view $i \in N$, the vector equality obtained from (3.7) by only keeping the pairs concerning this point of view, is exactly the multiset equality between $(\underbrace{y_i,\dots,y_i}_{n \text{ repetitions}}, a_i^{(1)},\dots,a_i^{(m)})$ and $(\underbrace{x_i,\dots,x_i}_{n \text{ repetitions}}, b_i^{(1)},\dots,b_i^{(m)})$. Therefore, the tuple E is an explanation of length m with n repetitions of the pair (x,y).

□

Example 3.1. Hotels are compared according to their comfort, offer of a restaurant, commute time and cost. We are given the *preference information* $\mathcal{P} \supseteq \{\pi_1,\pi_2,\pi_3\}$, with

π_1 :=((4* , no , 15 min , 180 $) , (2* , yes , 45 min , 50 $))
π_2 :=((4* , no , 45 min , 50 $) , (4* , yes , 15 min , 100 $))
π_3 :=((2* , yes , 15 min , 180 $) , (4* , no , 30 min , 180 $))

Explain why the alternative $x :=$ (4★, no, 45 min, 180 $) should be preferred to $y :=$ (2★, yes, 45 min, 100 $). We consider the dominance statement $d :=$ ((4★, yes, 30 min, 100 $), (4★, no, 30 min, 100 $))[c]. We claim that (π_1,π_2,d) is a syntactic cancellative explanation, of length three and without repetition, of the sought conclusion ((4★, no, 45 min, 180 $), (2★, yes, 45 min, 100 $)). The lists $\langle L_i\rangle_{i\in N}$ and $\langle R_i\rangle_{i\in N}$ built by Algorithm 3.1 are presented in Table 3.1, so that the syntactic check can be performed line by line.

Presenting the explanation. The depiction given in Table 3.1 seems quite technical and tedious, and might not be very enlightening. Therefore, we propose an alternative presentation, that we feel is better suited to convey

[b] In reference to the core, that satisfies (2.1). The augmented core is a useful demonstration tool hence, as it also encompasses the attributes of the positive query (x,y). The price to pay is that it is not static.

[c] As the two alternatives have similar attributes, except from the point of view of the presence of a restaurant, this statement is a complicated way of stating that, everything else being equal, the presence of a restaurant is preferable to the contrary. *Ceteris paribus statements* (see below) offer a more intuitive way of writing such statements.

	$\langle R_i \rangle_{i \in N}$			$\langle L_i \rangle_{i \in N}$	
Comfort:	2*	(from π_1)	$\sim_*$	2*	(from y)
	4*	(from π_2)	$\sim_*$	4*	(from π_1)
	4*	(from d)	$\sim_*$	4*	(from π_2)
	4*	(from x)	$\sim_*$	4*	(from d)
Restaurant:	no	(from d)	$\sim_r$	no	(from π_1)
	no	(from x)	$\sim_r$	no	(from π_2)
	yes	(from π_1)	$\sim_r$	yes	(from d)
	yes	(from π_2)	$\sim_r$	yes	(from y)
Commute time:	45 min	(from π_1)	$\sim_t$	45 min	(from π_2)
	45 min	(from x)	$\sim_t$	45 min	(from d)
	30 min	(from d)	$\sim_t$	30 min	(from y)
	15 min	(from π_2)	$\sim_t$	15 min	(from π_1)
Cost:	180 $	(from x)	$\sim_\$$	180 $	(from π_1)
	100 $	(from π_2)	$\sim_\$$	100 $	(from d)
	100 $	(from d)	$\sim_\$$	100 $	(from y)
	50 $	(from π_1)	$\sim_\$$	50 $	(from π_2)

Table 3.1: Verifying the syntactic cancellative explanation of Example 3.1.

meaningful information and, maybe, actual insight concerning the underlying reasoning, to the reader. Table 3.2 instantiates this presentation on the explanation given in Example 3.1—the same explanation that was presented in Table 3.1.

We know:	~~4*~~	no	~~15 min~~	180 $	π_1	2*	~~yes~~	45 min	~~50 $~~
	~~4*~~	~~no~~	45 min	~~50 $~~	π_2	~~4*~~	yes	~~15 min~~	~~100 $~~
	4*	~~yes~~	~~30 min~~	~~100 $~~	d	~~4*~~	~~no~~	~~30 min~~	100 $
Hence:	4*	no	45 min	180 $	$\mathcal{N}_{\{\pi_1,\pi_2\}}$	2*	yes	45 min	100 $

Table 3.2: Presentation template for syntactic cancellative explanations. Data is taken from Example 3.1.

In this presentation:

- each line of the premises is actually a unitary premiss, with a clear meaning—either mentioning an axiom of the preference information, 'you

told me that a is preferred to b', or stating the obvious 'a is preferred to b, as it is better on every aspect';

- the lines between premises and conclusion hint at an accrual of the atoms—pros on the left, cons on the right—into bundles;
- at the end of the premises block, it is implied that the bundle of pros outweighs the bundle of cons;
- cancellation carves the conclusion inside these bundles, in an 'obviously balanced' way, provided that it is legitimate to reason *ceteris paribus*.

Ceteris paribus statements. The presentation can be further simplified by integrating first-order cancellation directly into the scheme. We allow to provide *incomplete* preference statements as premises or conclusion, i.e. preference statements where identical values of the LHS and RHS are omitted and left blank. This use of syntactic sugar is illustrated by Table 3.3. For inline use, we denote $(x, y)_{N^{\neq}_{(x,y)}}$ such a partial statement, with an index precising the 'meaningful' points of view. Formally, such an incomplete statement represents a set of complete statements: all possible completions with similar attributes according to the points of view in $N \setminus N^{\neq}_{(x,y)}$. When the preference relation $\mathcal{R}$ is assumed to satisfy *first-order cancellation*,

$$\forall x, y \in \mathbb{X}, \quad (x, y)_{N^{\neq}_{(x,y)}} \subset \mathcal{R} \iff (x, y)_{N^{\neq}_{(x,y)}} \cap \mathcal{R} \neq \varnothing. \tag{3.9}$$

Example 3.2. The statements involved in the syntactic cancellative explanation of Example 3.1 can be written as follows:

$$\begin{aligned} \pi_2 &\equiv ((\text{no}, 45\text{ min}, 50\text{ \$}), (\text{yes}, 15\text{ min}, 100\text{ \$}))_{\{r,t,\$\}} \\ d &\equiv ((\text{yes}), (\text{no}))_{\{r\}} \\ (x, y) &\equiv ((4^*, \text{no}, 180\text{ \$}), (2^*, \text{yes}, 100\text{ \$}))_{\{*,r,\$\}} \end{aligned}$$

We know:	4*	no	~~15 min~~	180 $	π_1	2*	~~yes~~	~~45 min~~	~~50 $~~
		~~no~~	~~45 min~~	~~50 $~~	π_2		yes	~~15 min~~	100 $
		~~yes~~			d		~~no~~		
Hence:	4*	no		180 $	$\mathcal{N}_{\{\pi_1,\pi_2\}}$	2*	yes		100 $

Table 3.3: Syntactic cancellative explanation with ceteris paribus statements.

Cancellative flavor. The syntactic cancellative explanation pattern of length m looks a lot like the m-th order cancellation axiom (3.4). There is, nevertheless, a key difference, with the additional option of repeating the conclusion. This feature is unnecessary in some axiomatic settings, e.g.

- in a *value model*, when a statement of preference is represented by a statement about the sign of a difference in values; or
- in a *necessary and possible preference structure* [Giarlotta and Greco, 2013], if we suppose that both necessary and possible preference obey the cancellation axiom of any order: suppose we know that n repetitions of a statement of necessary preference are true, but not the original statement; thus, the inverse preference is possible, but in this case, the n repetitions of this possible statement becomes possible, which contradicts the premiss.

Nevertheless, if we want to make cancellation the sole inference engine for necessary preference, without reference to any other axiom besides compliance to dominance and preference information, the option of repeating the conclusion seems required.

Conjecture 3.3. *There is at least a context, with a set of criteria, alternatives, and a preference information $\mathcal{P}$ so that there is a pair in $\mathcal{N}_{\mathcal{P}}$ that cannot be explained by syntactic cancellative explanations without repetitions.*

3.1.3 Elliptic cancellation

In this section, we show that dominance statements are an unnecessary burden in the explanation, that can be alleviated by a modification of the rule governing cancellation. As the explanations are based on an omission (an *ellipsis*), we dub them *elliptic cancellative explanations*.

Focus on preference, not minutia. Instead of the accrual of arguments into bundles, the syntactic cancellative scheme can be seen as:

i) setting up a status quo, i.e.

$$y \oplus a^{(1)} \oplus \cdots \oplus a^{(m)} \equiv b^{(1)} \oplus \cdots \oplus b^{(m)} \oplus x$$

ii) removal of pros on both sides—$a^{(k)}$ from the LHS, $b^{(k)}$ from the RHS—until only y and x are left. When the value of each $a^{(k)}$ is known to be greater than the value of the corresponding $b^{(k)}$, it follows that the value left on the LHS, i.e. the value of y, is lesser than the value left on the RHS, i.e. the value of x.

The elliptic cancellative scheme we propose in this section simply replaces the first step by allowing the initial balance of value to begin skewed in favor of the RHS. This should only reinforce preference for x w.r.t. y, and eschews the painful and unnecessary requirement for striking a perfect balance at the initial step of the reasoning.

Definition 3.2 (elliptic cancellative explanation scheme)**.** *We say the tuple* $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathbb{X} \times \mathbb{X})^m$ *is an* elliptic cancellative explanation *of length* m *with* n *repetitions of the pair* (x, y) *if there exists a nonnegative integer* m' *and a* m'*-tuple of dominance statements* $\langle (c^{(1)}, d^{(1)}), \ldots, (c^{(m')}, d^{(m')}) \rangle \in \mathcal{D}^{m'}$ *such that* $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}, (c^{(1)}, d^{(1)}), \ldots, (c^{(m')}, d^{(m')}) \rangle$ *is a syntactic cancellative explanation of length* $m + m'$ *with* n *repetitions of the pair* (x, y)*.*

Example 3.3. Consider the following example, about the explanation of a query established as positive in Section 2.3.3. Assuming the same preference information as in Example 3.1, Equation (2.13) asserts that $(h_C, h_A) \in \mathcal{N}_{\{\pi_2, \pi_3\}}$. We consider the syntactic cancellative explanation given in Table 3.4 for this pair.

We know:		no	45 min	50 $	π_2		yes	15 min	100 $
	2*	yes	15 min		π_3	4*	no	30 min	
		yes	30 min	100 $	d_1		no	45 min	120 $
			20 min		d_2			35 min	
Hence:	2*	yes	20 min	50 $	$\mathcal{N}_{\{\pi_2, \pi_3\}}$	4*	no	35 min	120 $

Table 3.4: A syntactic cancellative explanation for Example 3.3.

The corresponding elliptic cancellative explanation is given in Table 3.5.

We know:		no	45 min	50 $	π_2		yes	15 min	100 $
	2*	yes	15 min		π_3	4*	no	30 min	
Hence:	2*	yes	20 min	50 $	$\mathcal{N}_{\{\pi_2, \pi_3\}}$	4*	no	35 min	120 $

Table 3.5: An elliptic cancellative explanation for Example 3.3.

The existence of an elliptic explanation is, by definition, tied to the existence of a syntactic explanation and is therefore a proof of necessary preference

Corollary 3.4. *Given some preference information $\mathcal{P} \subset \mathbb{X}^2$ and two alternatives $x, y \in X$, $(x, y) \in \mathcal{N}_{\mathcal{P}}$ if, and only if, there are two positive integers m and n and a tuple of pairs of alternatives $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in \mathcal{P}^m$ forming an elliptic explanation of length m with n repetitions of the pair (x, y).*

Verifying an elliptic explanation. The key idea with elliptic explanations is that the dominance information mentioned in syntactic explanations can be reconstructed on the fly and therefore does not need mentioning. Moreover, the verification of an elliptic cancellative explanation can completely bypass this reconstruction step, and be directly checked by Algorithm 3.2, a slightly modified version of Algorithm 3.1, used for checking syntactic cancellation. The only difference resides in the unitary check performed at the attribute level (line 7): while syntactic explanations are assessed through a syntactic check, elliptic explanations require a preference check, where the attributes appearing on the LHS—the one containing x—should be considered at least as good as the attributes appearing on the RHS—the one containing y.

Algorithm 3.2: Checking an elliptic cancellative explanation

Input:

- a pair of alternatives $(x, y) \in \mathbb{X}^2$;
- a tuple of pairs of alternatives $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathbb{X} \times \mathbb{X})^m$;
- a positive integer n.

Result: *True*, if the tuple $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle$ is an elliptic cancellative explanation with n repetitions of the pair (x, y); *False* else.

1 **foreach** *point of view* $i \in N$*:* **do**
2 **create** a list L_i of length $m + n$ containing the values $a_i^{(1)} \ldots a_i^{(m)}$ as well as n copies of y_i;
3 **sort** L_i in ascending order according to $\succsim_i$;
4 **create** a list R_i of length $m + n$ containing the values $b_i^{(1)} \ldots b_i^{(m)}$ as well as n copies of x_i;
5 **sort** R_i in ascending order according to $\succsim_i$;
6 **foreach** *index* $k \in [1..m + n]$ **do**
7 **if** $L_i[k] \prec_i R_i[k]$ **then**
8 **return** *False*;
9 **return** *True*.

Theorem 3.5. *Given a pair of alternatives* $(x, y) \in \mathbb{X}^2$*, two integers* m *and* n*, and a tuple of pairs of alternatives* $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathbb{X} \times \mathbb{X})^m$*, verifying if* $\langle (a^{(1)}, b^{(1)}), \ldots, (a^{(m)}, b^{(m)}) \rangle \in (\mathbb{X} \times \mathbb{X})^m$ *is an* elliptic cancellative explanation *of length* m *with* n *repetitions of the pair* (x, y) *can be performed in* $O(|N| \cdot (m+n) \ln(m+n))$ *operations by Algorithm 3.2.*

The proof of this theorem considers the iterative deletion of the attributes coming from dominance statements in the verification table. Before providing a formalization of this idea, we illustrate this principle on an example.

Example 3.4. (Example 3.3 continued) The verification table obtained when running Algorithm 3.1 on the syntactic cancellative explanation given in Table 3.4 is detailed in Table 3.6.

	$\langle R_i \rangle_{i \in N}$			$\langle L_i \rangle_{i \in N}$	
Comfort:	2*	(from x)	$\sim_*$	2*	(from π_3)
	4*	(from π_3)	$\sim_*$	4*	(from y)
Restaurant:	no	(from π_3)	$\sim_r$	no	(from π_2)
	no	*(from d_1)*	$\sim_r$	no	(from y)
	yes	(from π_2)	$\sim_r$	yes	(from π_3)
	yes	(from x)	$\sim_r$	*yes*	*(from d_1)*
Commute time:	*45 min*	*(from d_1)*	$\sim_t$	45 min	(from π_2)
	35 min	*(from d_2)*	$\sim_t$	35 min	(from y)
	30 min	(from π_3)	$\sim_t$	*30 min*	*(from d_1)*
	20 min	(from x)	$\sim_t$	*20 min*	*(from d_2)*
	15 min	(from π_2)	$\sim_t$	15 min	(from π_3)
Cost:	*120 \$*	*(from d_1)*	$\sim_\$$	120 \$	(from y)
	100 \$	(from π_2)	$\sim_\$$	*100 \$*	*(from d_1)*
	50 \$	(from x)	$\sim_\$$	50 \$	(from π_2)

Table 3.6: Checking the syntactic cancellative explanation of Example 3.1. Attributes of alternatives appearing in dominance statements are highlighted.

Deleting the dominance statements of the explanation from this verification table yields Table 3.7. This is also the verification table obtained when running Algorithm 3.2 on the elliptic cancellative explanation given in Table 3.5.

It can be observed that, while attribute values are now mismatched, the value appearing in the right column containing x are now always at least as strong, according to each point of view, as the one appearing in the left column containing y.

	$\langle R_i \rangle_{i \in N}$			$\langle L_i \rangle_{i \in N}$	
Comfort:	2*	(from x)	$\sim_*$	2*	(from π_3)
	4*	(from π_3)	$\sim_*$	4*	(from y)
Restaurant:	no	(from π_3)	$\sim_r$	no	(from π_2)
	yes	(from π_2)	$\succ_r$	no	(from y)
	yes	(from x)	$\sim_r$	yes	(from π_3)
Commute time:	30 min	(from π_3)	$\succ_t$	45 min	(from π_2)
	20 min	(from x)	$\succ_t$	35 min	(from y)
	15 min	(from π_2)	$\sim_t$	15 min	(from π_3)
Cost:	100 \$	(from π_2)	$\succ_\$$	120 \$	(from y)
	50 \$	(from x)	$\sim_\$$	50 \$	(from π_2)

Table 3.7: Checking the elliptic explanation presented in Table 3.5.

Proof. The two columns of the verification table are defined as follows:

- the list R_i contains the values of the tuple $(\underbrace{x_i, \ldots, x_i}_{n \text{ repetitions}}, b_i^{(1)}, \ldots, b_i^{(m)})$ sorted in ascending order according to $\succsim_i$; and
- the list L_i contains the values of the tuple $(\underbrace{y_i, \ldots, y_i}_{n \text{ repetitions}}, a_i^{(1)}, \ldots, a_i^{(m)})$ sorted in ascending order according to $\succsim_i$.

We prove that iteratively deleting dominance statements from a syntactic cancellative explanation of a pair (x, y) yields a verification table where, according to any point of view $i \in N$ the column containing x is pairwise-stronger than the column containing y. We call $R_i^{(j)}$ and $L_i^{(j)}$ the lists obtained from R_i and L_i respectively after j deletions.

i) Obviously, the lists $R_i^{(j)}$ and $L_i^{(j)}$ remain sorted.

ii) We prove by induction that, after any nonnegative number j of deletions, $R_i^{(j)}[k] \succsim_i L_i^{(j)}[k], \forall k \in [1..(m+n-j)]$.

- *Base case:* $j = 0$. By definition $R_i^{(0)} = R_i$ and $L_i^{(0)} = L_i$. As these original lists are obtained from a syntactic cancellative explanation, one is a permutation of the other. As they are sorted, they are, in fact, identical and therefore $R_i^{(0)}[k] \sim_i L_i^{(0)}[k],\ \forall k \in [1..(m+n)]$.
- *Inductive step:* suppose that, for a given number j of deletions, $R_i^{(j)}[k] \succsim_i L_i^{(j)}[k], \forall k \in [1..(m+n-j)]$. We consider the effect of the deletion of an additional dominance statement. From the point of view $i \in N$, this statement contributes to some value l_i indexed by

k_l to the list $L_i^{(j)}$, and to some value r_i indexed by k_r to the list $R_i^{(j)}$. As we consider a dominance statement, $l_i \succsim_i r_i$, and we can chose $k_l \geq k_r$ by virtue of the lists being sorted (see i) above). The lists obtained after deletion are given by:

$$L_i^{(j+1)}[k] = \begin{cases} L_i^{(j)}[k], & \text{if } k < k_l; \text{ or} \\ L_i^{(j)}[k+1], & \text{else.} \end{cases}$$

$$R_i^{(j+1)}[k] = \begin{cases} R_i^{(j)}[k], & \text{if } k < k_r; \text{ or} \\ R_i^{(j)}[k+1], & \text{else.} \end{cases}$$

Therefore,

$$(R_i^{(j+1)}[k], L_i^{(j+1)}[k]) = \begin{cases} (R_i^{(j)}[k], L_i^{(j)}[k]) & \text{if } k < k_r \\ (R_i^{(j)}[k+1], L_i^{(j)}[k]), & \text{if } k_r \leq k < k_l \\ (R_i^{(j)}[k+1], L_i^{(j)}[k+1]), & \text{if } k > k_l \end{cases}$$

In the second case, we note that $R_i^{(j)}[k+1] \succsim_i R_i^{(j)}[k]$ because the list $R_i^{(j)}$ is ordered in ascending order (see i) above). Therefore, in all three cases, $(R_i^{(j+1)}[k], L_i^{(j+1)}[k]) \in \succsim_i$ by induction hypothesis.

Reciprocally, any verification table where the sorted column $\langle R'_i \rangle$ containing x pairwise dominates the sorted column $\langle L'_i \rangle$ containing y can be expanded into a full syntactic explanation, by iterating insertion operations corresponding to dominance statements. If there is no point of view $i \in N$ such that the values of R'_i and L'_i differ, we have a syntactic cancellative explanation. Else, let $i \in N$ a point of view such that $L'_i \neq R'_i$, and k the first index where $R'_i[k] \succ_i L'_i[k]$. Then, inserting the *ceteris paribus* dominance statement $(R'_i[k], L'_i[k])_{\{i\}}$, everything being equal along every point of view in $N \setminus \{i\}$, into the explanation will insert the value $R'_i[k]$ into the L-column and the value $L'_i[k]$ into the R-column, changing one strict comparison into two equalities, and therefore yield an elliptic explanation with strictly less differing values between the two sets $(\underbrace{y_i, \ldots, y_i}_{n \text{ repetitions}}, a_i^{(1)}, \ldots, a_i^{(m)})$ and $(\underbrace{x_i, \ldots, x_i}_{n \text{ repetitions}}, b_i^{(1)}, \ldots, b_i^{(m)})$. □

3.1.4 Computing cancellative explanations

This section is devoted to finding explanations for a given pair of alternatives in the necessary preference relation. We begin by establishing a connection between the elliptic cancellative explanation defined in the previous section and tools presented in Chapter 2. This leads to the diagram represented

in Figure 3.1, connecting the notion of necessary preference to its various explanations. Then, we address the problem of finding short explanations: we prove it is NP-complete, and we provide an elegant mixed integer programming formulation permitting to solve it.

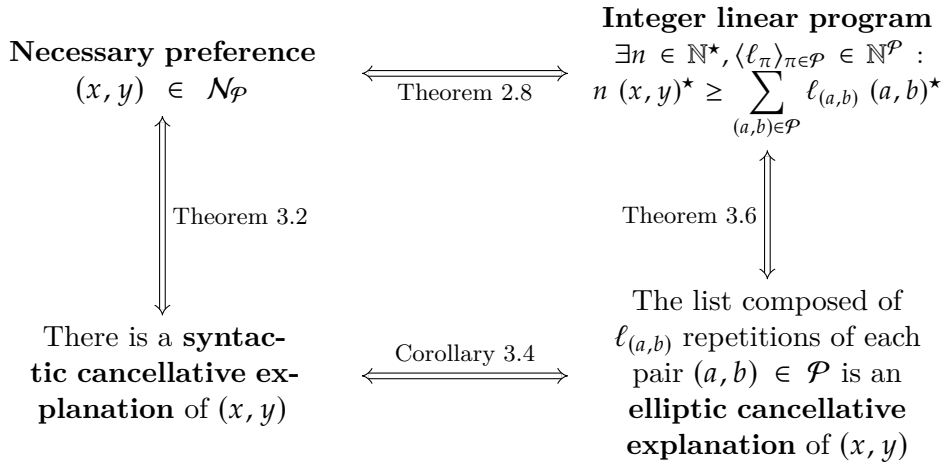

Figure 3.1: Diagram of characterizations of necessary preference, for a pair of alternatives that is not unbounded, with references to the properties.

The following theorem completes the diagram of Figure 3.1.

Theorem 3.6. *For any pair of alternatives (x, y), for any positive integer n, for any tuple of pairs of alternatives E, we define*

$$\widehat{\mathbb{P}}_i := \bigcup_{(a,b)\in\mathcal{P}} \{a_i\} \cup \bigcup_{(a,b)\in\mathcal{P}} \{b_i\} \cup \{x_i\} \cup \{y_i\}; \tag{3.10}$$

If the pair (x, y) is not left unbounded by E, the following propositions are equivalent:

(i) E is an elliptic cancellative explanation with n repetitions of the pair (x, y);

(ii) for all nondecreasing value functions $\langle v_i \rangle_{i\in N} \in \prod_{i\in N} \left((\widehat{\mathbb{P}}_i, \succsim_i) \to (\mathbb{R}, \geq) \right)$,

$$n \cdot \sum_{i\in N} \big(v_i(x_i) - v_i(y_i) \big) \geq \sum_{i\in N} \sum_{(a,b)\in E} \big(v_i(a_i) - v_i(b_i) \big)\,; \tag{3.11}$$

(iii) with any core containing $\prod_{i\in N} \widehat{\mathbb{P}}_i$,

$$n \cdot (x, y)^\star \geq \sum_{(a,b)\in E} (a, b)^\star;$$

(iv) with any core containing $\prod_{i\in N}\left(\bigcup_{(a,b)\in\mathcal{P}}\{a_i\}\cup\bigcup_{(a,b)\in\mathcal{P}}\{b_i\}\right)$,

$$n\cdot(x,y)^\star \geq \sum_{(a,b)\in E}(a,b)^\star;$$

Proof.

- (i) $\iff$ (ii)
  Remark that Equation (3.11) is equivalent to:

$$\sum_{i\in N}\left(n\cdot v_i(x_i)+\sum_{(a,b)\in E}v_i(b_i)\right)\geq\sum_{i\in N}\left(n\cdot v_i(y_i)+\sum_{(a,b)\in E}v_i(a_i)\right) \quad (3.12)$$

  Projection of this comparison into the double-dual of each $\widehat{\mathbb{P}}_i$ corresponds exactly to the preference check performed by Algorithm 3.2, and is therefore equivalent to E being an elliptic explanation of (x,y) according to Theorem 3.5.

- (ii) $\iff$ (iii)
  We recall that, for any point of view $i\in N$ and alternatives z^1,z^2 such that the attribute values z_i^1,z_i^2 belong to the core $\mathbb{P}_i$, the following identity holds:

$$\sum_{k=1}^{|\mathbb{P}_i|-1}(z^1,z^2)_{i,k}^\star\cdot\left(v_i(p_i^{k+1})-v_i(p_i^k)\right)=v_i(z_i^1)-v_i(z_i^2) \quad (3.13)$$

  Therefore, (iii) is merely (ii) expressed in the base $\left\langle\left(v_i(p_i^{k+1})-v_i(p_i^k)\right)\right\rangle_{(i,k)}$.

- (iv) $\Rightarrow$ (ii)
  When alternatives x and y do not belong to the core, Equation (3.13) is no longer satisfied. Because of the *pessimistic rounding* (as explained in Section 6, Component 2.3.2), the linear form $\sum_{k=1}^{|\mathbb{P}_i|-1}(x,y)_{i,k}^\star\cdot\left(v_i(p_i^{k+1})-v_i(p_i^k)\right)$ actually underestimates the value of $v_i(x_i)-v_i(y_i)$.

- (iii) $\Rightarrow$ (iv)
  Suppose $n\cdot(x,y)^\star\geq\sum_{(a,b)\in E}(a,b)^\star$ holds on the 'augmented' core $\mathbb{P}_i$. Removing the value(s) of $\mathbb{P}_i\cap\left(\bigcup_{(a,b)\in\mathcal{P}}\{a_i\}\cup\bigcup_{(a,b)\in\mathcal{P}}\{b_i\}\right)$ creates a coarser scale, where each interval of the coarse scale is represented by a coefficient that is equal to the coefficient of some interval of the fine scale. This property, valid as soon as the pair (x,y) is not unbounded by the core, is illustrated by Figure 3.2. Therefore, the comparison assumed for the fine scale transfers to the coarse scale.

□

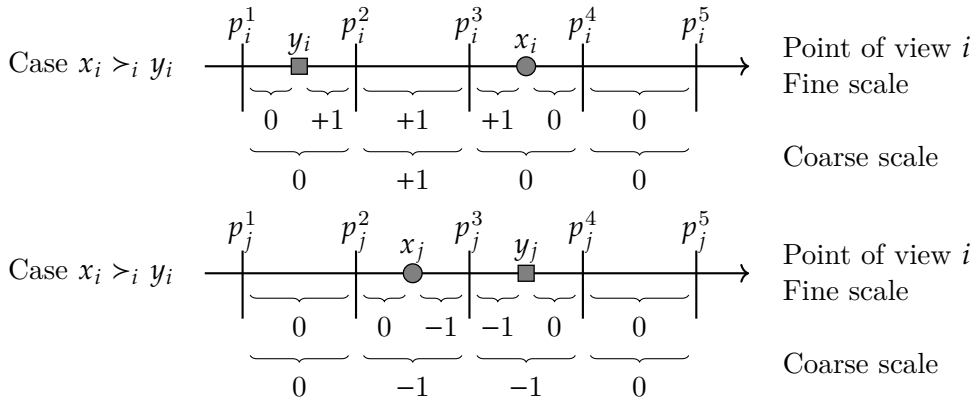

Figure 3.2: Effect of subdividing the core intervals to account for the attribute values of x and y on the covector $(x, y)^\star$. Subdividing yields greater values, but there remains an interval of the finer subdivision that has the same orientation as the coarser interval. Cases where the attributes of x or y fall outside the range of the core intervals, or fall inside the same core interval, are not represented.

Interpretation. Theorem 3.6 highlights the relationship between the theoretical and computational aspects of ncessary preference under the assumption of additive values, studied in Chapter 2, and their explanation. Both cancellative explanation schemes introduced in this section are simply vehicles for the fundamental relation

$$n \cdot (x, y)^\star \geq \sum_{(a,b)\in E} (a, b)^\star;$$

This relation makes clear the particular way of assessing trade-offs associated with the additive value model:

- points of view are assessed independently—visually, each one is associated to an axis;
- a trade-off is represented by $|N|$ one-dimensional vectors corresponding to the displacement from an initial alternative to a final one;
- a trade-off is *beneficial* when it corresponds to an increase in value, i.e. the final alternative is preferred to the initial one;
- each statement of the preference information is an assertion of benefit concerning some trade-off;
- a dominance trade-off is obviously beneficial;
- Proposition (iii) asserts that, according to each point of view, n times the displacement from y to x is, according to each point of view, at least as

beneficial as the sum—as one-dimensional vectors—of the displacements representing the premises;

- in addition, Proposition (iv) asserts the irrelevance of the exact position of the endpoints of the vectors representing the displacement from y to x.

Figure 3.3 illustrates this geometric interpretation of a cancellative explanation.

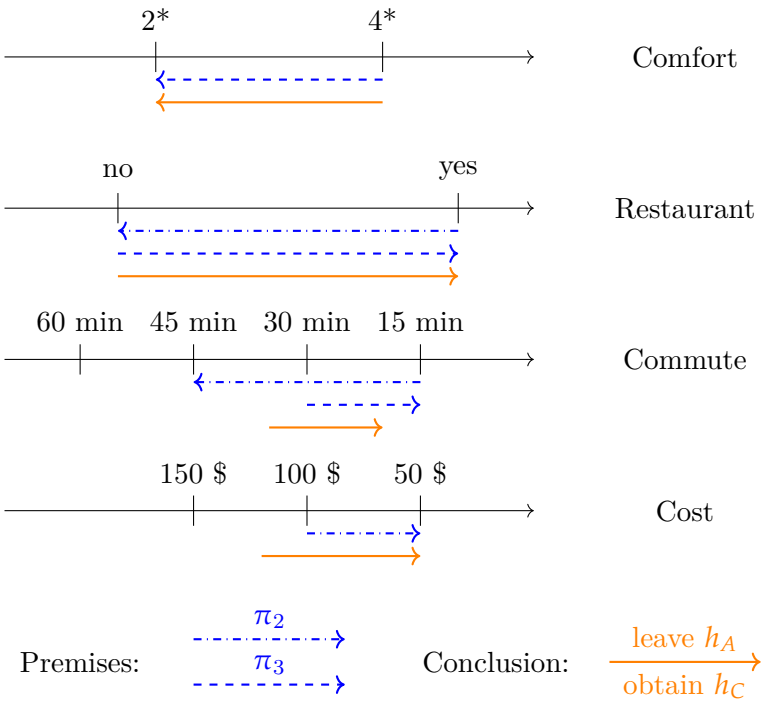

Figure 3.3: The elliptic cancellative explanation for Example 3.3 represented in vector form.

Simple explanations. It seems reasonable to believe that an explanation is easier to process by a cognitive agent—'simpler'—when it is short. In the case of cancellative explanations, the actual cognitive burden mainly comes from three factors: the number of points of view $|N|$, that we consider as mostly exogenous; the length m of the premises; and the number n repetitions of the conclusion. Without any experimental evidence, we consider the problem of finding an explanation for a given pair $(x, y) \in \mathcal{N}_{\mathcal{P}}$ which is as simple as

possible as a bi-objective integer linear minimization problem:

$$\min_{n,m\in\mathbb{N}^\star}(n,m) \quad \text{such that} \quad \begin{cases} n\ (x,y)^\star \geq \sum_{\pi\in\mathcal{P}} \ell_\pi \pi^\star; & \text{and} \\ m \geq \sum_{\pi\in\mathcal{P}} \ell_\pi. \end{cases} \tag{3.14}$$

Integer linear programs offer a powerful language permitting to describe difficult combinatorial problems. These formulations can be given wholesale to dedicated solvers, that eschews the need for developing a dedicated piece of software and benefits from state-of-the-art refinements in the solving of such problems. Nevertheless, it would be unwise to delegate the search for a short explanation of a given pair of alternatives to such a solver, if this search were not, intrinsically, a difficult combinatorial problem. The following theorem addresses this issue.

Theorem 3.7. *The problem of deciding, for a given input* $(x,y,n,m) \in \mathbb{X}\times\mathbb{X}\times\mathbb{N}^\star\times\mathbb{N}^\star$ *if there is an elliptic cancellative explanation of the pair* (x,y) *of length at most* m *with at most* n *repetitions is NP-complete.*

Proof. Theorem 3.5 proves an elliptic cancellative explanation can be verified in polynomial time. Therefore, this decision problem belongs to the class NP.

Hardness can be established e.g. by reduction from Vertex cover [Karp, 1972, Cormen et al., 2001]. Formally, a vertex cover V' of an undirected graph $G=(V,E)$ is a subset of V such that $uv\in E \Rightarrow u\in V' \vee v\in V'$, that is to say it is a set of vertices V' where every edge has at least one endpoint in the vertex cover V'. The Vertex cover problem consists in, given an instance (G,k) where $G=(V,E)$ is a graph and k a positive integer, to *decide* whether G has a vertex cover of size at most k, or not. Given an instance of Vertex cover, we map it to a gadget instance of our problem:

- the set of points of view is $N = V\cup E$;
- an alternative is a subset of N;
- each point of view is evaluated on a binary scale, with presence preferred to absence;
- we define the preference information as containing all statements of the form $(\{(u,v)\},\{u,v\})$—any edge is preferred to the set of its endpoints—for all edges $(u,v)\in E$.

Any elliptic cancellative explanation without repetition of the pair (E,V)—the pros are the edges, the cons are the vertices—of length k is a subset of E that forms a vertex cover of size k of of the graph G, and reciprocally.

□

Conclusion

We introduced the notion of *cancellative explanations*, based on the accrual of premises to obtain a conclusion. We studied this explanative framework in the light of the principles stated in Section 1.3.4.

Completeness. Every preference statement that can be skeptically inferred from the preference information and the way of reasoning corresponding to the additive value model is supported by a cancellative explanation.

Soundness. Every preference statement that is supported by a cancellative explanation can be skeptically inferred from the preference information and the way of reasoning corresponding to the additive value model;

Simplicity. We provided several ways of presenting cancellative explanations, in the form of tables, diagrams, or argument schemes, and proposed to ground them on a syntactic check, or alternatively to keep implicit the information tied to dominance, which can easily be restored by the recipient, in the spirit of *enthymemes*. We provided polynomial-time algorithms to verify the explanations. We proposed a metric to evaluate the simplicity of these explanations, and formulated the problem of finding explanations as simple as possible.

Computation. Remarkably, while adjudicating necessary preference is a polynomial problem, explaining it concisely is NP-complete.

3.2 Explaining via a sequence of preference swaps

One might regret the cancellative explanations detailed in Section 3.1 are not well suited to narration. Precisely, considering a cancellative explanation with m premises, for each point of view, the accrual of the premises is mathematically represented by a sum of m vector coordinates. While using the vector sum as a m-ary operator is perfectly legitimate, thanks to the associativity of the binary sum, it might overwhelm the capacity of the recipient of the explanation to make sense of it. Information is presented in a parallel form, which does not correspond to a narrative structure, sequential by nature.

In this section, we explore an alternate venue, reminiscent of the *even-swaps* method [Hammond et al., 1998], that aims at building a sequence of simple arguments to support a given pairwise preference. The section unfolds as follows: in Section 3.2.1, we recall the functioning of the *even-swaps* method; in

Section 3.2.2, we detail our explanation engine; in Section 3.2.3 we question this engine in the light of our expectations concerning the support for accountable decision making; in Section 3.2.4 we provide positive results in the case where we try to explain the necessary preference relation inferred from preference information obtained on a binary core.

3.2.1 Sequences of *even-swaps*

The *even-swaps* method [Hammond et al., 1998] is an interactive and constructive elicitation procedure assuming conditions close to the ones ensuring an additive value model of preferences[d]. This method aims at identifying, between two options x and y, which one is preferred to the other, without explicitly constructing the utility functions. This is basically an iterative elimination process based on trade-offs between pairs of attributes ("swaps"), that can be seen as a scattered exploration of the level set containing x of the value function representing the preferences of the decision maker.

Starting from $e^{(0)} := x$, the aim is, at each step j, to construct an alternative $e^{(j+1)}$ such that either:

a) $e^{(j+1)} \equiv y$, if either $(e^{(j)}, y) \in \mathcal{D}$ or $(y, e^{(j)}) \in \mathcal{D}$; or

b) $e^{(j+1)} \sim e^{(j)}$ and $\exists i_1 \neq i_2 \in N$ such that $\forall i \in N \setminus \{i_1, i_2\},\ e^{(j+1)} \sim_i e^{(j)}$.

A method propitious to explanation. Considered through the prism of explanation, even-swaps have several very attractive features :

- the method produces a sequence of exchanges, which is naturally suited to being narrated;
- each exchange step between an alternative and the next in the chain is purposely simple to grasp, as the transaction either occurs along solely two points of view, or is a dominance statement;
- the elements referenced by the method are entirely situated inside the decision space $\mathbb{X}$; no artifact of the underlying decision model (such as utility functions), or relations between criteria, are ever put forward;
- as opposed to the cancellative explanation schemes defined in Section 3.1, the inference from premises to conclusion is not parallel; instead, each explanatory step merges two premises into a conclusion that serves as a premiss in the next explanatory step; *transitivity* serves as an implicit basis for reasoning;

[d] More precisely, the *even-swaps* method only requires solvability w.r.t. a single criterion. A discussion on the solvability conditions ensuring an additive representation can be found in e.g. [Gonzales, 2003].

The chain of arguments implicitly put forward during the elicitation permitting to adjudicate a query $(x, y) \in \mathbb{X}^2$ can be made explicit through the explanation schemes presented in Table 3.8, for the successive indifference statements, and Table 3.9 for the final step.

We know:	• x is equally preferred to $e^{(j)}$; and • $e^{(j)}$ is equally preferred to $e^{(j+1)}$.
Therefore,	x is equally preferred to $e^{(j+1)}$.

Table 3.8: Indifference scheme for an explanation based on *even-swaps*.

- Case where x is preferred to y:

We know:	• x is equally preferred to $e^{(j)}$; and • $e^{(j)}$ dominates y.
Therefore,	x is preferred to y.

- Case where y is preferred to x:

We know:	• y dominates $e^{(j)}$; and • $e^{(j)}$ is equally preferred to x.
Therefore,	y is preferred to x.

Table 3.9: Dominance schemes for an explanation based on a sequence of swaps.

A method based on indifference statements. At each step but the last, the *even-swaps* method constructs the following alternative $e^{(j+1)}$ by selecting two points of view $i_1 \neq i_2 \in N$, and an attribute level $\ell \in \mathbb{X}_{i_2}$—usually $\ell \equiv y_{i_2}$, in order to ensure the convergence of the process—then asking to the decision maker the following question:

"Which level of satisfaction, according to the point of view i_1, would compensate for a change from $e^{(j)}_{i_2}$ to ℓ according to the point of view i_2, everything else being equal [to the situation described by $e^{(j)}$][e] ?" Such a questioning procedure severely limits the practicality of the *even-swaps* approach.

[e]The part between brackets does not need to be mentioned if we assume the preferences follow the *first-order cancellation* axiom (see Section 3.1.1), that allows to reason *ceteris paribus*.

- Indifference requires *compensation* between criteria [Krantz et al., 1971]—the assumption that, no matter how fit an alternative is according to some point of view, it can always be surpassed overall by improving sufficiently according to any other single point of view, barring the possibility that some difference in attributes on one criterion could be impossible to compensate for.
- Indifference requires *solvability* of the attribute scales [Krantz et al., 1971]—the assumption that, when two alternatives that are similar according to all points of view but one compare differently to a same third one, there is a satisfaction level of this particular criterion in between that permits to strike indifference with the third alternative. Solvability naturally occurs on continuous scales but rarely between discrete ones.
- Indifference imposes a high cognitive workload on the decision maker, as it repeatedly requires the procurement of a very precise information.
- Indifference is hardly a robust notion, especially in the context of incomplete preferences[f].

Consequently, in the next section, we propose a generalization of even-swaps that avoids these issues, while retaining their simplicity and being well suited to the context of incomplete preference.

Swaps as *trade-offs*. The active elicitation of an additive value model in multiple criteria decision aiding is not the only domain where the notion of *swap* is used. For instance, in Software Engineering, the Architecture Tradeoff Analysis Method (ATAM) is used in order to assess software architectures according to "quality attribute goals" [Kazman et al., 2000]. A *trade-off point* is an architecture parameter affecting at least two quality attributes in different directions. For example, increasing the speed of the communication channel improves throughput in the system but reduces its reliability. Thus the speed of that channel is a trade-off point. The concept of trade-off point in ATAM makes explicit the interdependencies between attributes. Even though trade-offs can be defined for any number of attributes, the examples of trade-offs that are provided by experts are almost always given on pairs of attributes. This is the case of the example provided above.

[f]We note that [Mustakoji and Hamalainen, 2005, Mustakoji and Hamalainen, 2007] also propose to enrich the original even swaps method in a way that accounts for incomplete knowledge about the value function. They consider a "practical dominance" notion when the value of an alternative is at least as high as the value of another one with every feasible combination of parameters, this perspective being very close to the one developed in [[Greco et al., 2008]. However, this notion is only used for pre-processing dominated alternatives, and not integrated in the swap process, let alone used for explanatory purposes.

3.2.2 Explaining with a sequence of *preference-swaps*

We introduce the notion of *preference-swaps*, as a tool for the explanation of a binary relation $\mathcal{R} \subseteq \mathbb{X}^2$ satisfying:

i) Pareto: $\mathcal{D} \subseteq \mathcal{R}$;

ii) transitivity; and

iii) optionally, *first-order cancellation*.

Obviously, these requirements are fulfilled by any preference relation supported by an additive value model, and also, given some preference information $\mathcal{P} \subset \mathbb{X}^2$ by the *necessary preference relation assuming an additive value model* $\mathcal{N}_{\mathcal{P}}$.

In *preference swaps*, compared to *even-swaps* [Hammond et al., 1998], we relax both the notion of indifference between consecutive alternatives in the sequence $e^{(0)} := x, e^{(1)}, \ldots, e^{(n)} := y$ and the notion of elementary swap:

- Indifference is replaced by an assumption of (weak) preference : $(e^{(j-1)}, e^{(j)}) \in \mathcal{R}$.
- In the context of explanation, we propose to use the syntactic complexity of a pair $(x, y) \in \mathbb{X}^2$—its *order*—as a proxy for the cognitive complexity of evaluating the acceptability of the transaction of giving y to obtain x. Dominance relations are deemed to be simple, and are given the lowest order. For relations requiring trade-offs between criteria, we define the order of a swap as the number of differing attributes between the two alternatives. This assumption is all the more reasonable as the preference relation $\mathcal{R}$ has the property of *first-order cancellation*, which allows to omit the shared attributes in any statement.

Definition 3.3 (preference swaps of order k)**.** *Given a binary relation $\mathcal{R}$ on $\mathbb{X}$ satisfying* Pareto*, for all positive integer k, the set of preference-swaps of order k is the binary relation $\mathcal{R}^k \subseteq \mathbb{X} \times \mathbb{X}$ defined by:*

$$\mathcal{R}^{(k)} := \begin{cases} \mathcal{D}, & \text{if } k = 1; \\ \{(x, y) \in \mathcal{R} \setminus \mathcal{D}, |N^{\neq}_{(x,y)}| = k\}, & \text{if } k > 1. \end{cases} \tag{3.15}$$

This definition permits to break down the relation $\mathcal{R}$ into layers of inherent syntactic complexity: $\mathcal{R} = \bigcup_{1 \leq k \leq |N|} \mathcal{R}^{(k)}$.

We can now define the notion of explanation by a sequence of preference swaps. This type of explanation breaks down a single preference statement $(x, y) \in \mathcal{R}$ that the decision maker needs to understand into a sequence of several preference statements $(e^{(j-1)}, e^{(j)}) \in \mathcal{R}$. The idea is that the initial preference (x, y) is complex to understand as the values of x and y differ on most (if not all) attributes, whereas each intermediate comparison $(e^{(j-1)}, e^{(j)})$

is much easier to understand as it involves alternatives differing according to only a few points of view.

Definition 3.4 (Explanation by preference swaps, order and length)**.** $\forall (x, y) \in \mathbb{X}^2, n \in \mathbb{N}$, *an explanation of length* n *of the pair* (x, y) *for the relation* $\mathcal{R}$ *is a tuple* $(e^{(0)}, e^{(1)}, \ldots, e^{(n)}) \in \mathbb{X}^{n+1}$ *such that* $e^{(0)} = x, e^{(n)} = y$ *and* $\forall j \in \mathbb{N} : 1 \leq j \leq n \ , \ (e^{(j-1)}, e^{(j)}) \in \mathcal{R}$. *The order of such explanation is the integer* $\hat{k} = \max\{k \in \mathbb{N} : \exists (j \in \mathbb{N} : 1 \leq j \leq n) \ , \ (e^{(j-1)}, e^{(j)}) \in \mathcal{R}^{(k)}\}$.

3.2.3 Properties of the *preference-swaps* explanation engine

In this section, we are interested in qualifying the soundness and completeness of the system of explanations based on preference swaps, as well as the simplicity of its artifacts.

Soundness. When $\mathcal{R}$ is transitive, an explanation of a pair of alternatives is a *proof* that this pair belongs to $\mathcal{R}$. Hence, explanations via sequences of preference swaps form a *sound* explanatory system.

Finding a balance between completeness and simplicity. One can note that somehow we have two elements to appreciate the quality of the explanation. First, the number of comparisons (swaps) used to construct such an explanation. Second, its complexity which is defined by the most complex or difficult swap (with the highest order).

However, an important question regarding a pair $(x, y) \in \mathbb{X}^2$ is whether it is possible to find an explanation by preference-swaps of the pair (x, y). The answer obviously depends on the bound, if any, placed upon the order of the swaps linking the explanation chain, or the length of the explanation chain. We address this issue by first putting a cap on the order (the order of an explanation being the order of its most difficult link), then looking for the possibility of finding an explanation subject to this order constraint. Then, if explanations are available, we look for short ones.

Definition 3.5 (pairs explainable by low-order preference swaps)**.** *For any positive integer* k, $\mathcal{E}_k(\mathcal{R})$ *is the set of pairs* $(x, y) \in \mathbb{X}^2$ *for which there exists an explanation with a sequence of any length of preference swaps of order at most* k.

There is a trade-off between the value of the cap placed upon the order of explanations and the set of pairs we are able to explain.

Theorem 3.8.

$$\mathcal{D} = \mathcal{E}_1(\mathcal{R}) \subseteq \mathcal{E}_2(\mathcal{R}) \subseteq \cdots \subseteq \mathcal{E}_k(\mathcal{R}) \subseteq \cdots \subseteq \mathcal{E}_{|N|}(\mathcal{R}) = \mathcal{R}$$

We interpret this nesting as a hierarchy of complexity inside the relation $\mathcal{R}$, that goes beyond the mere syntactic complexity directly measured by the order.

Proof.

- for any $(x, y) \in \mathcal{E}_1(\mathcal{R})$, there is a tuple $(e^{(0)}, e^{(1)}, \ldots, e^{(n)}) \in \mathbb{X}^n$ such that $e^0 = x, e^{(n)} = y$ and $\forall j \in \mathbb{N} : 1 \leq j \leq n$, $(e^{(j-1)}, e^{(j)}) \in \mathcal{D}$. As relation $\mathcal{D}$ is transitive, $(x, y) \in \mathcal{D}$, hence $\mathcal{D} \supseteq \mathcal{E}_1(\mathcal{R})$. Conversely, the sequence $(e^{(0)} := x, e^{(1)} := y)$ is an explanation of length one and of order one of any pair $(x, y) \in \mathcal{D}$, hence $\mathcal{D} \subseteq \mathcal{E}_1(\mathcal{R})$. Finally, $\mathcal{D} = \mathcal{E}_1(\mathcal{R})$;
- for $k' \geq k$, an explanation of order at most k is also an explanation of order at most k', so $\mathcal{E}_k(\mathcal{R}) \subseteq \mathcal{E}_{k'}(\mathcal{R})$;
- the sequence $(e^{(0)} := x, e^{(1)} := y)$ is an explanation of length one and of order $|N^{\neq}_{(x,y)}|$ of any pair $(x, y) \in \mathcal{R}$. As $|N^{\neq}_{(x,y)}| \leq |N|$, $\mathcal{R} \subseteq \mathcal{E}_{|N|}(\mathcal{R})$. Conversely, an explanation (of any order and any length) of a pair (x, y) is a proof by transitivity of $(x, y) \in \mathcal{R}$, thus $\mathcal{R} \supseteq \mathcal{E}_{|N|}(\mathcal{R})$. Finally, $\mathcal{R} = \mathcal{E}_{|N|}(\mathcal{R})$

□

Computational aspects. When a cap $k_{\max}$ is placed upon the order of acceptable swaps, the problems of checking the explainability of a given statement $(x, y) \in \mathcal{R}$ and, if positive, of finding the shortest explanation, can be formally described using the directed graph $\mathcal{G} := (\mathbb{X}, \bigcup_{1 \leq k \leq k_{\max}} \mathcal{R}^{(k)})$.

- *Checking explainability:* The pair (x, y) belongs to $\mathcal{E}_{k_{\max}}(\mathcal{R})$ if, and only if, there is an explanation of the pair (x, y) via a sequence of preference swaps of order at most $k_{\max}$ if, and only if, the vertices x and y are *connected* in the graph $\mathcal{G}$.
- *Computing the simplest explanation:* if x and y are connected in $\mathcal{G}$, the shortest explanation for the pair (x, y) is the *shortest path* between x and y in $\mathcal{G}$. Thus, the length of a shortest explanation is bounded by the *diameter* of $\mathcal{G}$.

Of course, there are efficient algorithms in order to decide if a graph is connected or not [Even and Tarjan, 1975], to compute the shortest path between two

vertices of a graph, or to compute the diameter of a graph [Aingworth et al., 1996]. However, it may be challenging to use them with regard to the size—possibly infinite, and, when finite, exponential in the number of criteria—of the graph $\mathcal{G}$.

Existence of arbitrarily long shortest explanations. Keeping the explanation short has a great bearing on its ability to convince. Even if each elementary comparison $(e^{(j-1)}, e^{(j)}) \in \mathcal{R}$ is trivial for the decision maker, the overall sequence $(x, e^{(1)}, \ldots, e^{(n-1)}, y)$ cannot be seen as a convincing explanation if it is too long. One then looks for the shortest possible explanations, and hope for an upper bound on this minimal size. Unfortunately, as soon as there are three criteria measured on infinite scales, this diameter has no upper bound, as expressed by the following theorem.

Theorem 3.9. *For any integer p, if there is a subset $A \subseteq N : |A| = 3$ and $\forall i \in A,\ |\mathbb{X}_i| \geq p$, then there is a relation $\mathcal{R}$ satisfying Pareto, transitivity and* first-order cancellation*, and a pair $(x, y) \in \mathcal{R}^{(3)}$ such that $(x, y) \in \mathcal{E}_2(\mathcal{R})$ and any explanation of (x, y) by preference swaps of order at most 2 has a length greater than $2p$.*

Proof. The proof requires instantiating the relation $\mathcal{R}$, and is presented in Appendix A. We make use of the necessary preference relation under the assumption of additive values, for some carefully built preference information. □

3.2.4 Results in the case of a binary core

This section is devoted to the presentation of positive results concerning the use of the explanation engine powered by preference-swaps, under specific assumptions:

1. the relation to be explained is the necessary preference relation, assuming additive values;
2. the preference information only reference two levels according to each point of view;
3. the explanations are limited to chaining preference-swaps of order at most two, i.e. dominance statements and trade-offs along two points of view.

To account for the core composed of binary scales, we define:

$$\forall i \in N,\ \mathbb{B}_i := \{\top_i \succ_i \bot_i\}, \quad \mathbb{B} := \prod_{i \in N} \mathbb{B}_i \tag{3.16}$$

The restriction to binary scales allows us to introduce a simpler notation for preference, in terms of positive or negative arguments:

Definition 3.6 (pros and cons of a necessary preference statement)**.** *If* $\mathcal{P} \subset \mathbb{B}^2, \forall (x, y) \in \mathcal{N}_{\mathcal{P}}$,

$$\begin{aligned}(x,y)^+ :=& \quad \{i \in N : (x,y)^{\star}_{(i,1)} = +1\} &= \{i \in N : y_i \precsim_i \bot_i \prec_i \top_i \precsim_i x_i\}\\ (x,y)^- :=& \quad \{i \in N : (x,y)^{\star}_{(i,1)} = -1\} &= \{i \in N : \bot_i \precsim_i x_i \prec_i y_i \precsim_i \top_i\}\end{aligned}$$

Example 3.5 (Example 3.3 continued)**.** We are still interested in comparing hotels according to the points of view of comfort, the presence of a restaurant, the commute time to the city center, and the cost. We consider the binary core described by Table 3.10.

i	Point of view	$\bot_i$	$\top_i$
*	Cost	2*	4*
r	Restaurant	no	yes
t	Commute	40 min	20 min
$	Cost	100 $	50 $

Table 3.10: Binary core attribute values.

$\rho_1 :=((\top_*, \bot_r, \top_t, \bot_\$), (\bot_*, \top_r, \bot_t, \top_\$))$, so $\rho_1^+ = \{*, t\}$ and $\rho_1^- = \{r, \$\}$;
$\rho_2 :=((\bot_*, \bot_r, \bot_t, \top_\$), (\bot_*, \top_r, \top_t, \bot_\$))$, so $\rho_2^+ = \{\$\}$ and $\rho_2^- = \{r, t\}$;
$\rho_3 :=((\bot_*, \top_r, \top_t, \top_\$), (\top_*, \bot_r, \bot_t, \top_\$))$, so $\rho_3^+ = \{r, t\}$ and $\rho_3^- = \{*\}$.

We note that:

- ρ_1 is weaker than π_1;
- ρ_2 and π_2 are two different completions of the same ceteris paribus statement;
- ρ_3 is weaker than π_3.

Thus, we expect $\mathcal{N}_{\{\rho_1,\rho_2,\rho_3\}} \subset \mathcal{N}_{\{\pi_1,\pi_2,\pi_3\}}$.

Structure of an explanation We are interested in characterizing $\mathcal{E}_2(\mathcal{N}_{\mathcal{P}})$ when $\mathcal{P} \subset \mathbb{B}^2$. Assuming binary reference scales, the relation $\mathcal{N}^{(2)}_{\mathcal{P}} \subset \mathbb{X}^2$ between alternatives induces a relation between points of view $\widetilde{\mathcal{N}}^{(2)}_{\mathcal{P}} \subset N^2$: when, everything else being equal, being 'good' according to some point of view i and 'bad' according to another point of view i' is preferred to the converse, we add the edge (i, i') to the graph of the relation $\widetilde{\mathcal{N}}^{(2)}_{\mathcal{P}}$.

Definition 3.7 (swaps between points of view)**.** *If* $\mathcal{P} \subset \mathbb{B}^2$*,*

$$\widetilde{\mathcal{N}}_{\mathcal{P}}^{(2)} := \{(i, i') \in N^2 : i \neq i' \text{ and } ((\top_i, \bot_{i'}), (\bot_i, \top_{i'}))_{\{i,i'\}} \subset \mathcal{N}_{\mathcal{P}}^{(2)}\}$$

We note that building the graph of $\widetilde{\mathcal{N}}_{\mathcal{P}}^{(2)}$ is polynomial time w.r.t. the number of points of view $|N|$ and the number of statements in $\mathcal{P}$, as there are $|N|^2$ potential edges, each requiring to solve a linear program in $|N|$ dimensions with $|\mathcal{P}|$ constraints (see Theorem 2.8, characterization n°3 of necessary preference).

Example 3.6 (Example 3.5 continued)**.** Assuming additive values, the necessary preference relation deduced from the preference information given in Example 3.5 induces the following criteria swaps:

$$\widetilde{\mathcal{N}}_{\mathcal{P}}^{(2)} = \{(*, r), (t, r), (\$, *), (\$, r), (\$, t)\}$$

.

The graph of this relation is represented on Figure 3.4. The nodes are labeled so as to make explicit the semantic of each edge.

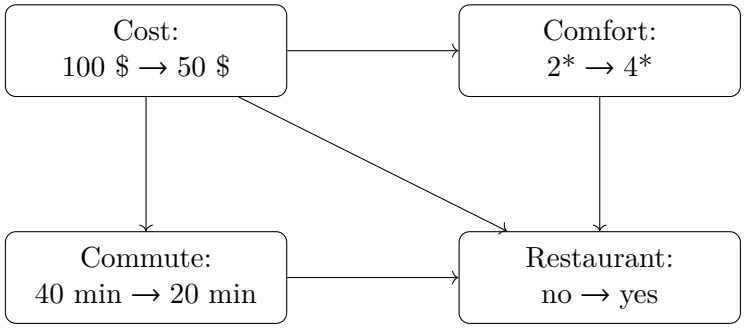

Figure 3.4: Binary relation between criteria.

For instance, the compact criteria swap statement $(\$, r)$, represented by the arrow from \$ to r, indicates $((\text{no}, 50\ \$), (\text{yes}, 100\ \$))_{\{r,\$\}} \subset \mathcal{N}_{\{\rho_1,\rho_2,\rho_3\}}$, i.e. everything else being equal, the decision maker prefers a cheap hotel without a restaurant than an expensive hotel with a restaurant. It can also be interpreted in terms of *intensity of preference*[g]: the preference for a reduction in cost from

[g]The quaternary relation between alternatives at the heart of the notion of *intensity of preference* is discussed in [Bouyssou and Pirlot, 2004, Bana e Costa and Vansnick, 1994, Bana e Costa and Vansnick, 1995]. [Figueira et al., 2009] proposes a framework for the robust elicitation of an additive value model where preference information can be given in the form of intensity of preference statements.

100 $ to 50 $ is more intense than the preference for an amelioration from the point of view of the presence of a restaurant.

We emphasize though that this relation is not suitable to being presented *directly* as an explanation. This graph could easily be interpreted erroneously as a relation of comparative importance between points of view, regardless of the attribute values. While this notion is relevant for some preference models, such as those based on a lexicographical order of the points of view [Fishburn, 1976, Wilson and George, 2017], it would be fallacious here. In our additive context, each edge of the graph corresponds to a speech act and to a potential transition $e^{(j)} \to e^{(j+1)}$ in an explanation via a sequence of preference-swaps[h]. Also, the acts represented by the edges of the graph are *atomic*, because the core is binary, there is no point in considering intermediate values, i.e. splitting an edge. Because we restrict ourselves to preference swaps of order at most two, there is also no point in considering relations between the nodes of the graph with an arity greater than two. The following theorem reveals how the graph of $\widetilde{\mathcal{N}}_{\mathcal{P}}^{(2)}$ permits to decide on the explainability of a pair of alternatives, to build an explanation when it is possible, and highlights the structure of explanations.

Theorem 3.10 (Term by term explanation)**.**
If $\mathcal{P} \subset \mathbb{B}^2, \forall \sigma \in \mathcal{N}_{\mathcal{P}}$, the following propositions are equivalent :

1. *Explainability with a sequence of preference swaps of order at most two:* $\sigma \in \mathcal{E}_2(\mathcal{N}_{\mathcal{P}})$
2. *Integral combination of statements:*
   $\exists a \in \mathbb{N}^\star, \gamma_1, \cdots, \gamma_q \in \widetilde{\mathcal{N}}_{\mathcal{P}}^{(2)}, \ell_1, \cdots, \ell_q \in \mathbb{N}, m_1, \cdots, m_n \in \mathbb{N}$:
   $$a\sigma^\star = \sum_k \ell_k \gamma_k^\star + \sum_k m_k \delta_{(k,1)}^\star$$
3. *Reduction to* Maximum bipartite matching*:*
   There is a matching of cardinality $|\sigma^-|$ in the graph of $\widetilde{\mathcal{N}}_{\mathcal{P}}^{(2)} \cap (\sigma^+ \times \sigma^-)$.
4. *Term-by-term explanation:*
   There is an injection $\phi : \sigma^- \to \sigma^+$ such that $\forall k \in \sigma^-, (\phi(k), k) \in \widetilde{\mathcal{N}}_{\mathcal{P}}^{(2)}$.

Proof. See Appendix A. □

[h]This contrasts with the usual abstract argument framework [Dung, 1995], where the arguments are the nodes of a graph and can be involved in several attacks simultaneously. In such an abstract framework, there is no notion of *state space* and *current state*, whereas these notions are crucial for explanations based on a sequence of alternatives.

In a nutshell, an explanation is a sequence where, at each step, a positive argument is used up to cancel an inferior negative argument, and, eventually, every negative argument has been cancelled. We highlight three consequences of this theorem.

Completeness: *If the preference information only refers to swaps of order two, then every necessary preference statement can be explained by swaps of order at most two.*

Corollary 3.11 (case of 2-order preference statements)**.** *If* $\mathcal{P} \subset \mathbb{B}^2$*, and* $\forall \pi \in \mathcal{P}, |N_\pi^{\neq}| = 2$ *then* $\mathcal{N}_\mathcal{P} = \mathcal{E}_2(\mathcal{N}_\mathcal{P})$*.*

Proof. By theorem 3.8, $\mathcal{E}_2(\mathcal{N}_\mathcal{P}) \subset \mathcal{N}_\mathcal{P}$. Reciprocally, if $(x, y) \in \mathcal{N}_\mathcal{P}$, the implication 1. $\Rightarrow$ 4. of theorem 2.8 ensures the existence of a linear combination with integral, non-negative coefficients $n(x, y)^\star = \sum_{\pi \in \mathcal{P}} \ell_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} m_{(i,k)} \delta^\star_{(i,k)}$. The assumption that $\forall \pi \in \mathcal{P}, |N_\pi^{\neq}| = 2$ entails $\mathcal{P} \subset \widetilde{\mathcal{N}}^{(2)}_\mathcal{P}$, so this linear combination satisfies proposition 2 of theorem 3.10, thus $(x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$ by proposition 1. □

Simplicity: *Explanations can be kept short.*
The next corollary states that, in the favorable case of preference information expressed on a binary core, the length of an explanation by a sequence of preference swaps of order at most two is at most "half the number of points of view, rounded down, plus one", which appears manageable for the recipient of explanation.

Corollary 3.12 (short explanations)**.** *If* $\mathcal{P} \subset \mathbb{B}^2$*, for any statement* $(x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$*, there exists an explanation with a length at most* $\lfloor \frac{|N|}{2} \rfloor + 1$*, where* $\lfloor m \rfloor$ *denotes the integer part of* m*.*

The bound $\lfloor \frac{|N|}{2} \rfloor + 1$ basically come from the fact that $|(x, y)^-| \leq \lfloor \frac{|N|}{2} \rfloor$, which follows directly from the characterization n°4 of Theorem 3.10. The main asset of this theorem is that it is constructive. The explanation sequence will be provided in the next section.

Computation: *Building an explanation, or ensuring there is none, is handled by an efficient algorithm.*
Indeed, it boils down to building the bipartite graph of the relation $\widetilde{\mathcal{N}}^{(2)}_\mathcal{P} \cap (\sigma^+ \times \sigma^-)$, then solve the polynomial-time Maximum bipartite matching problem on this instance. If the matching returned has cardinality $|\sigma^-|$, it can be interpreted in the sense of a *term-by-term explanation*, where every point

of view against preference is counterbalanced by a point of view in favor of preference. We emphasize that in this context, arguments, carried by edges, are *one-shot*—the same argument is never used twice.

Example 3.7 (Example 3.6 continued)**.** We consider the preference information $\{\rho_1, \rho_2, \rho_3\}$, and we look for explanations for preference statements:

- *Explain why* (2*, no, 20 min, 50 $) *should be preferred to* (4*, yes, 40 min, 120 $)?

  Denoting $\sigma := ((2*, \text{no}, 20\text{ min}, 50\ \$), (4*, \text{yes}, 40\text{ min}, 120\ \$))$ we have $\sigma^+ = \{t, \$\}$ and $\sigma^- = \{*, r\}$.
  Figure 3.5 represents the graph of the relation $\widetilde{\mathcal{N}}^{(2)}_{\{\rho_1,\rho_2,\rho_3\}} \cap (\sigma^+ \times \sigma^-)$, and highlights a matching of cardinality two in this graph: $\phi_\sigma^{-1} = \{(\$, *), (t, r)\}$.

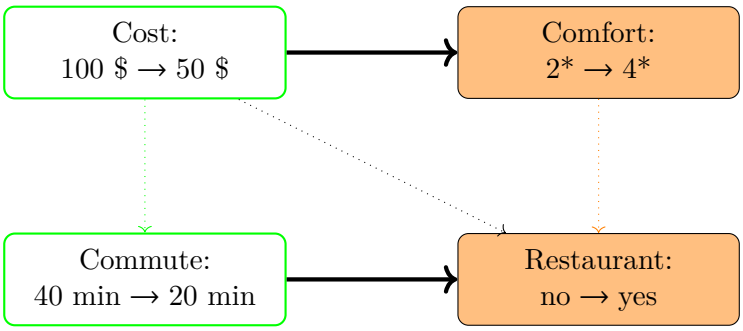

Figure 3.5: Matching cons (in solid orange boxes) with stronger pros (in green frames).

  This matching can be interpreted according to the following explanation:

  ($, *): *"The gain in cost (from 100 $ to 50 $) is preferable to the loss in comfort (from 4* to 2*)"*;

  (t, r): *"The gain in commute time (from 40 min to 20 min) is preferable to the loss of a restaurant"*.

  This explanation is parallel by nature: the two lines of argumentation do not interfere, as they concern disjoint sets of points of view. They can therefore be presented in any order. Each possible permutation leads to a different sequence of preference-swaps:

  – either ((2*, no, 20 min, 50 $) $\xrightarrow{(\$,*)}$ (4*, no, 20 min, 120 $) $\xrightarrow{(t,r)}$ (4*, yes, 40 min, 120 $));

  - or ((2*, no, 20 min, 50 \$) $\xrightarrow{(t,r)}$ (2*, yes, 40 min, 50 \$) $\xrightarrow{(\$,*)}$ (4*, yes, 40 min, 120 \$)).

- *Explain why* (3*, no, 20 min, 50 \$) *should be preferred to* (3*, yes, 40 min, 120 \$)?
  Actually, this statement is true, because it corresponds to the preference information conveyed by ρ_2, strengthened by dominance. Nevertheless, this statement cannot be supported by a sequence of preference-swaps of order at most two, because it is solely supported along the point of view of cost, that can not simultaneously match the disadvantages from the points of view of the presence of a restaurant and the commute time.

Part II

Sorting with a Noncompensatory model

Foreword

We give a brief definition of the topics discussed in this Part—sorting problems and noncompensatory preference models—and a correspondingly brief sample of decision aiding situations where they can be relevant. We give an overview of the contents of each chapter, and list the published papers of this author that serve as a basis for this text.

Sorting into ordered categories

Throughout this part, we focus on *sorting* problems, where we are interested on adjudicating the absolute fitness of any alternative, described according to conflicting points of view, on the basis of its own merits. The outcome of this adjudication is a *category*, where:

- the number of categories is *finite*—say p;
- the categories are known in advance: $\{C^1, \ldots, C^p\}$, as opposed to clustering problems where categories emerge from data;
- categories are ordered by *level of requirement*: $\{C^1 \prec \cdots \prec C^p\}$, corresponding to the aggregated preference.

Noncompensatory models of preferences

In this part we assume that preferences are expressed through a *noncompensatory* filter. For models interested in the comparative assessment of alternatives, [Fishburn, 1976] proposes a filter that solely encodes the nature of the points of view according to which an alternative is preferred to another; [Bouyssou, 1986] proposes to relax this assumption to account for possible *discordance* effects, where a very large difference in fitness according to a point

of view has a 'veto' effect and weakens a strict preference into an indifference judgment. For models yielding an absolute evaluation of alternatives, e.g. value models in general, or models specifically tailored for sorting, the lack of compensation is reflected by a cardinality restriction on the codomains of the marginal value functions. In the strictest acceptation, [Bouyssou and Marchant, 2007a, Bouyssou and Marchant, 2007b] requires that, for each level of requirement, each point of view is expressed and accounted for through a binary language: an alternative is either worthy, or not, barring the possibility of compensating e.g. unworthy attributes with 'very worthy' ones.

Applications of Noncompensatory sorting

While the assumption of an absence of compensation may appear drastic, simplistic or abusive, there are many real-life decision situations that are well addressed by this model.

Binary preferences

- *Intrinsically binary values:* in some situations, a point of view may be naturally measured by a 2-valued criterion: given a system, a spare part—a motherboard for a computer, a tyre for a car, a kidney for a transplant—is either compatible, or not[a].
- *Approximation of value:* 2-valued criteria may offer a passable approximation of a sigmoid fitness.

Binary representations. Noncompensation can be intrinsically tied to preference, but can also stem from its chosen representation.

- *Organizational constraint:* For instance, in voting context, it is customary to limit the expressiveness of a ballot, sometimes down to a single bit per alternative such as in a referendum or in *approval voting* [Laslier and Sanver, 2010]. We can think this limitation as favoring *clarity*—of the stakes of the vote as well as of the adjudication process.
- *Interpretability:* In some Machine Learning applications, especially for the medical domain [Sokolovska et al., 2017], models are constrained in their language in order to be readable and computable by humans, without any assisting device. This built-in demand leads de facto to noncompensatory models.

[a] this is obviously a simplification of reality

Reference levels. Noncompensatory models for sorting are intrinsically tied to the notion of *reference levels*: levels of attributes between between which the model is blind to difference in fitness, because of the re-encoding of the scales. These levels offer a metaphor that can serve as a useful base for interaction during a decision aiding process, leveraged in e.g. the Electre method [Roy, 1991], or in the domain *specification engineering*: in order to assess candidate technological solutions to a given problem, it might be appropriate to compare them, on every aspect, to the existing situation, or to an explicit target. Reference level may also serve in a descriptive approach: see e.g. the Reference-Dependent Theory [Tversky and Kahneman, 1991], where a single level describes the perceived current situation and serves as a reference for assessing changes.

Research questions

This part is devoted to address the following issue with the state of the art concerning the use of noncompensatory models for decision aiding situations modeled after a sorting problem:

Research question II.1. *Many sorting models are based on a noncompensatory description of preference. How do these approaches relate to each other, or differ from each other?*

Research question II.2. *Until now, indirect approaches to the elicitation of noncompensatory sorting models based on mathematical programming suffer from poor computational efficiency, that restrict them to solving toy instances. Is it possible to do better?*

Research question II.3. *How an efficient solver dedicated to the feasibility of the inverse noncompensatory sorting problem can be called upon and built around in order to enhance the accountability of decision aiding processes?*

Chapters

The chapters composing this part address the research questions related to noncompensatory sorting in an array of manners:

- Chapter 4 proposes an answer to the Question II.1, by installing the *noncompensatory sorting model* (NCS) by [Bouyssou and Marchant, 2007a, Bouyssou and Marchant, 2007b] in a central role. The chapter illustrates the functioning of this model on a working example, proposes a taxonomy of variants found in the literature, and highlights some

shortcomings of the state of the art concerning the elicitation of its parameter.

- Chapter 5 is deliberately dry and technical, and devoted to the formal aspects of solving the inverse NCS problem (Inv-NCS), consisting in deciding whether a given assignment of reference alternatives to ordered categories can be represented in the noncompensatory sorting model and, in such a case, finding adequate parameters of the model. It partly addresses Question II.2 by exploring an alternate approach to this problem and providing two Boolean satisfiability formulations for Inv-NCS.
- Chapter 6 details some experimental studies aimed at assessing the viability of the formulations detailed in the previous chapter, compared to the state-of-the-art elicitation methods for noncompensatory sorting based on mixed integer programming. The results obtained contribute to address Question II.2.
- Chapter 7 proposes some clues regarding the manners to put the Inv-NCS machinery to use in decision aiding contexts, with demands in terms of accountability. It therefore addresses Question II.3.

Featured contributions

This part results from the augmentation, harmonization, and the putting into perspective of several published work by this author and others:

- [Belahcene et al., 2018c] introduces the first formulation presented in Chapter 5, and the corresponding experimental study presented in Chapter 6;
- [Belahcene et al., 2018a] introduces the second formulation presented in Chapter 5, and sets up some of the issues discussed in Chapter 7;

See also

Of course, the action of answering questions entails the asking of new ones. The research agenda for this part is already dense. In order to provide the reader with a fairly streamlined narration, we omitted to provide additional contributions, that we judged too tangential, or requiring further development, to get incorporated to the main material. This supplementary material appears in the appendices at the end of the book. We give here a brief overview of their content.

Noncompensatory sorting without frontiers

Additional research question II.4. *The noncompensatory sorting model NCS is based on limiting profiles. Is it possible to circumvent this notion?*

This question has already received attention, in e.g. [Fishburn, 1976]. We detail an attempt at proposing a model for noncompensatory sorting based on comparative judgments in Appendix B, reproducing [Belahcene et al., 2017b].

Comparing the SAT formulations for Inv-NCS

Additional research question II.5. *Chapter 5 details two formulations based on Boolean satisfiability in order to solve the inverse noncompensatory sorting problem, but Chapter 6 only investigates the performance of one of them. How does the pairwise formulation behave experimentally? How do the two formulations compare to each other?*

[Belahcene et al., 2018b] addresses this issue, and is reproduced in Appendix C.

A SAT formulation for noncompensatory ranking with RMP

Additional research question II.6. *The noncompensatory approach to modeling preference is not reserved to sorting problems. Indeed, the* ranking with multiple reference points *model (RMP) applies the same notions to ranking problems, where preference is binary by nature. Could the contributions to the description of the inverse problem for NCS, using Boolean satisfiability formulations, be transposed, mutatis mutandis, to RMP?*

[Belahcene et al., 2018d] explores this venues, with promising results reproduced in Appendix D.

4

The noncompensatory sorting model

4.1 The noncompensatory sorting model (NCS)

This section is devoted to the presentation of the noncompensatory sorting model, introduced in [Bouyssou and Marchant, 2007a, Bouyssou and Marchant, 2007b]. Following the authors, Section 4.1.1 introduces the case of two categories separated by a single boundary profile, and Section 4.1.2 formally addresses the general case, with any number of categories. Section 4.1.3 offers a straightforward interpretation of a specific aggregation procedure following the noncompensatory sorting model, with a simple semantic for the parameters. It also proposes some insight about the requirements made on the parameters. Section 4.2 proposes a working example, that can serve as a tutorial for decision aiding with this model. Section 4.3 tries to situate the NCS model on the atlas of sorting models, identifying meaningful variants and establishing some relations with other identified models.

4.1.1 Sorting with a single profile.

In the noncompensatory sorting model NCS, profiles define the boundaries between categories. Therefore, a single profile corresponds to the case where alternatives are sorted between two ordered categories that we label, without loss of generality as Good and Bad. A specific aggregation procedure is described by a pair of parameters:

- a limiting profile $b \equiv \langle b_i \rangle_{i \in N}$ that defines, according to each point of view $i \in N$, an upper set $\mathcal{A}_i$ of approved values at least as good as b_i (and, by contrast, a lower set $\mathbb{X} \setminus \mathcal{A}_i$ of disapproved values strictly worse than b_i); and
- a set $\mathcal{T}$ of sufficient coalitions of points of view, required to be an upset of the power set of the points of view.

These notions are combined into the following assignment rule:

$$\forall x \in X, \quad x \in \textsc{Good} \iff \{i \in N : x \succsim_i b\} \in \mathcal{T}$$

An alternative is considered as Good if, and only if, it is better than the limiting profile b according to a sufficient coalition of points of view.

4.1.2 Sorting into multiple categories

With p categories, the parameter space is extended accordingly, with approved sets $\langle \mathcal{A}_i^k \rangle_{i \in N,\ k \in [2..p]}$ and sufficient coalitions $\langle \mathcal{T}^k \rangle_{k \in [2..p]}$ declined per boundary. The ordering of the categories $\{C^1 \prec \cdots \prec C^p\}$ translates into a nesting of the sufficient coalitions:

$$\forall k \in [2..p],\ \mathcal{T}^k \text{ is an upset of } (2^N, \subseteq) \text{ and } \mathcal{T}^2 \supseteq \cdots \supseteq \mathcal{T}^p \tag{4.1}$$

and also a nesting of the approved sets:

$$\forall i \in N,\ \forall k \in [2..p],\ \mathcal{A}_i^k \text{ is an upset of } (\mathbb{X}_i, \precsim_i) \text{ and } \mathcal{A}_i^2 \supseteq \cdots \supseteq \mathcal{A}_i^p \tag{4.2}$$

These tuples of parameters are augmented on both ends with trivial values:

$$\mathcal{T}^1 := \mathcal{P}(N),\ T^{p+1} := \varnothing, \quad \forall i \in N,\ \mathcal{A}_i^1 = \mathbb{X},\ \mathcal{A}_i^{p+1} = \varnothing, \tag{4.3}$$

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

$$NCS_\omega(x) = C^k \iff \begin{cases} \{i \in N : x \in \mathcal{A}_i^k\} & \in \mathcal{T}^k \\ \text{and } \{i \in N : x \in \mathcal{A}_i^{k+1}\} & \notin \mathcal{T}^{k+1}. \end{cases} \tag{4.4}$$

4.1.3 Interpretation of the aggregation procedure

The generic model, with p categories, is a bit less straightforward to interpret than the model with two categories and a single profile. We propose the following narrative interpretation:

- at each *level of requirement* $k \in [1..p]$, an alternative is deemed good enough if, and only if, it is approved at exigence level k for a coalition of points of view that is sufficient at the level of requirement k;
- at a given level of requirement, the judgment according each point of view is 2-valued. Extremely bad and mildly bad values are treated the same, and so are extremely good values w.r.t. mildly good values, barring compensatory trade-offs—situations where excellence according to a single point of view, as opposed to mere fitness, would be allowed to trump slight mediocrity according to several points of view;
- the nesting of the approved sets and sufficient coalitions means that approval becomes more stringent the higher the level of requirement goes. Particularly, it prevents judgment reversals—situation where an alternative could be considered good enough at some level k but not at some lower level $k' < k$—and allows the $p-1$ queries ⟨ 'is the candidate good enough at level k'⟩$_{k\in[1..p-1]}$ to be adjudicated in any order without modifying the final assignment of the candidate[a]. Indeed, a stronger form of the noncompensatory sorting rule can be given:

$$NCS_\omega(x) = C^k \iff \begin{cases} \forall k' \leq k, & \{i \in N : x \in \mathcal{A}_i^{k'}\} \in \mathcal{T}^{k'}; \text{ and} \\ \forall k' > k, & \{i \in N : x \in \mathcal{A}_i^{k'}\} \notin \mathcal{T}^{k'}. \end{cases} \tag{4.5}$$

4.2 A working example

Terry is a journalist and prepares a car review for a special issue. He considers a number of popular car models, and wants to sort them in order to present a sample of cars "selected for you by the editorial board" to the readers.

This selection is based on 4 criteria : cost (measured in dollars), acceleration (measured by the time, in seconds, to reach 100 km/h from full stop – lower is better), braking power and road holding, both measured on a qualitative scale ranging from 1 (lowest performance) to 4 (best performance). The performances of six models are described in Table 4.1.

In order to assign these models to a class among $C^{1^\star}$ (average) $\prec C^{2^\star}$ (good) $\prec C^{3^\star}$ (excellent), Terry considers 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 4.2. The resulting labeling of the six alternatives according to each criterion is depicted in Figure 4.1 and Table 4.3.

[a] This situation should be put in perspective with the fixed order in which the reference points are examined in the RMP model.

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 4.1: Performance table.

profile	cost	acceleration	braking	road holding
$b^{1\star}$	17 250	30	2.2	1.9
$b^{2\star}$	15 500	28.8	2.5	2.6

Table 4.2: Limiting profiles.

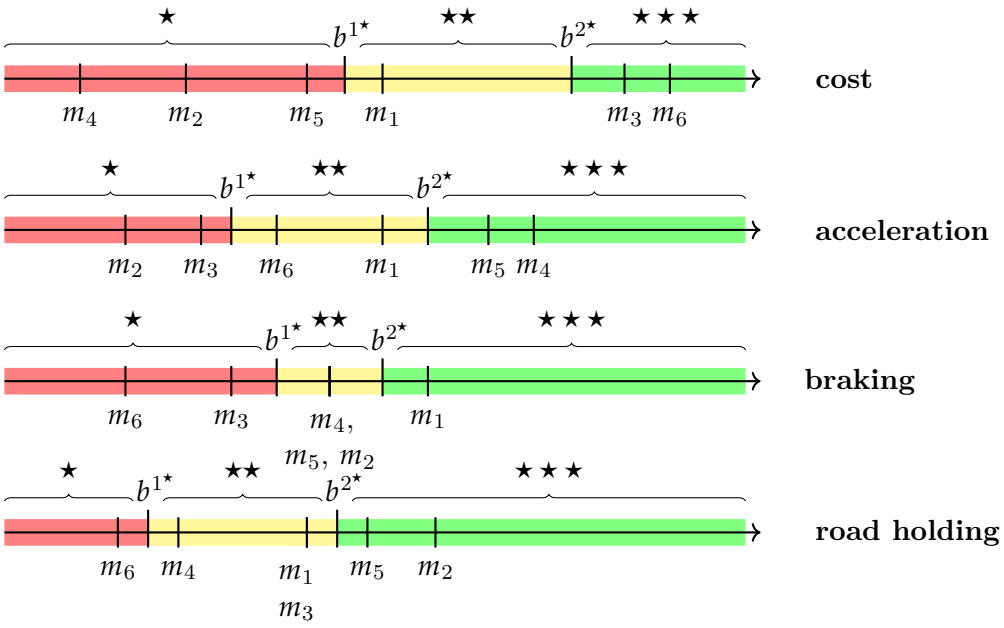

Figure 4.1: Representation of performances w.r.t. category limits.

- 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*

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

Table 4.3: Categorization of performances.

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 4.2.

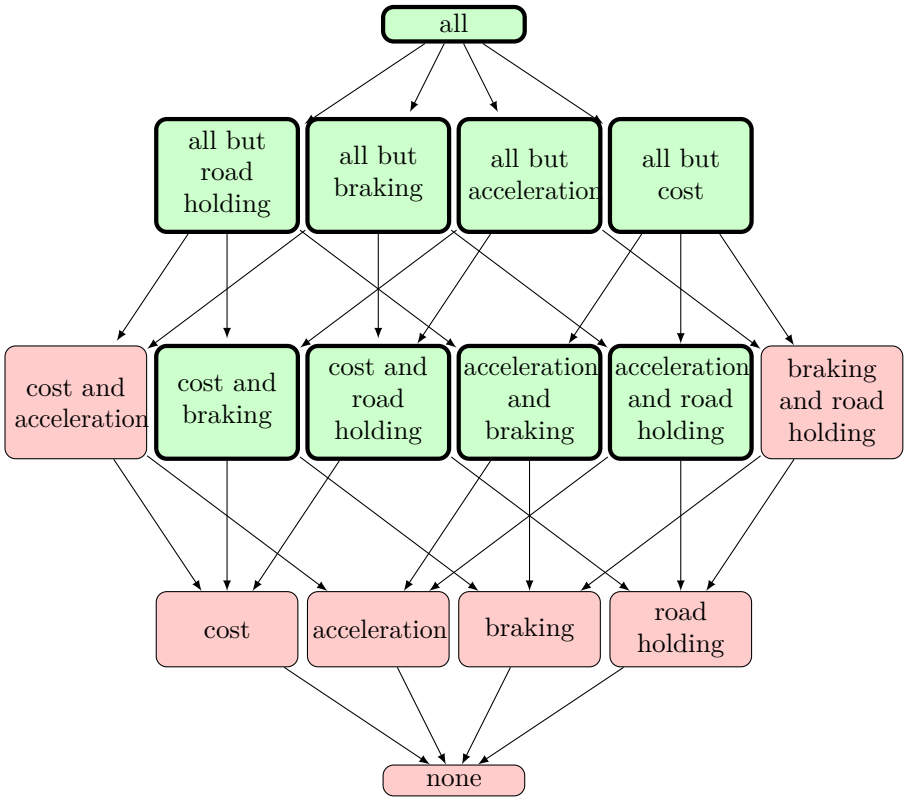

Figure 4.2: Sufficient (green-filled and thick-bordered) and insufficient (red-filled and thin-bordered) coalitions of criteria. Arrows denote strength—pointing towards the weaker.

Finally, the model yields the following assignments (Table 4.4):

model	assignment
m_1	★★
m_2	★
m_3	★★
m_4	★★
m_5	★★★
m_6	★

Table 4.4: Model Assignments.

4.3 Variants

In this section, we mention a number of variants of the noncompensatory sorting model that can be found in the literature. Some of these variants are simplifications of the model, with additional assumptions that restrict the parameters—limiting profiles and sufficient coalitions—either explicitly or implicitly. Other variants extend the model, either by removing some assumptions or by augmenting the expressive power of the model with additional parameters.

4.3.1 Presentation with profiles

In the most popular implementations of the noncompensatory sorting model—the ancestor Electre-Tri[b] [Mousseau et al., 2000, Roy, 1991] and its recent, streamlined avatar MR-Sort [Leroy et al., 2011]—there is no mention of 'approved sets'. These variants instead rely on the notion of *limiting profiles*, that act as kind of frontiers between categories. This difference in presentation is purely cosmetic

Approved sets and boundary profiles are tied by the following axiom:

$$\forall i \in N,\ k \in [2..p],\ x_i \in \mathbb{X}_i, \quad x_i \in \mathcal{A}_i^k \iff x_i \succsim_i b_i^k$$

Condition (4.2) translates into an ordering of the values $\langle\ b_i^k\ \rangle_{k\in[2..p]}$ according to $\succsim_i$ for a given criterion $i \in N$, and an ordering of the boundary profiles according to dominance $\mathcal{D}$:

$$b^2, \ldots, b^p \text{ is a non-decreasing sequence of } (\mathbb{X}, \mathcal{D})$$

This sequence is also conveniently augmented by an ideal profile b^{p+1} and an anti-ideal profile b^1.

[b] itself derived from the ranking model Electre III [Roy, 1978].

The noncompensatory sorting rule can be written with parameters $\langle\, b \,\rangle$ and $\langle \mathcal{T} \rangle$:

$$NCS_{(\langle\, b \,\rangle, \langle \mathcal{T} \rangle)}(x) = C^k \iff \begin{cases} \{i \in N : x \succsim_i b_i^k\} & \in \mathcal{T}^k; \text{ and} \\ \{i \in N : x \succsim_i b_i^{k+1}\} & \notin \mathcal{T}^{k+1}. \end{cases}$$

This equivalent manner of presenting the model highlights the link between the noncompensatory models NCS for sorting and RMP (Ranking with Multiple Profiles) [Rolland, 2013] for ranking.

The idea of representing an upset by means of a threshold can be extended. When dealing with totally ordered sets, such as the $\langle (\mathbb{X}, \succsim_i) \rangle_{i \in N}$, this threshold boils down to a single value b_i. For partially ordered sets, such as $(\mathcal{P}(N), \subseteq)$, this threshold takes the form of an *antichain*—a subset where any two elements are incomparable. Therefore, sufficient coalitions are sometimes represented as upper closures of an antichain of *minimal winning coalitions*[c] $\mathcal{Z}$, i.e. $B \in \mathcal{T} \iff \exists Z \in \mathcal{Z} : Z \subset B$. Unfortunately, this representation is not more compact than an explicit representation of the power set, as the length of an antichain of the hypercube can be exponentially long.

4.3.2 Limited array of parameters.

The set of preference parameters – all the pairs $(\langle b \rangle, \langle \mathcal{T} \rangle)$ satisfying (4.1) and (4.3.1) – can be considered too wide and too unwieldy for practical use in the context of a decision aiding process. Therefore, following [Bouyssou and Marchant, 2007b], one may consider to explicitly restrict either the sequence of limiting profiles, or the sequence of sufficient coalitions:

- *noncompensatory sorting with a unique set of sufficient coaltions*

$$\mathcal{T}^2 = \cdots = \mathcal{T}^p; \tag{4.6}$$

- *noncompensatory sorting with a unique limiting profile*

$$b^2 = \cdots = b^p. \tag{4.7}$$

  Or, equivalently,

$$\forall i \in N,\ \mathcal{A}_i^2 = \cdots = \mathcal{A}_i^p. \tag{4.8}$$

4.3.3 k-additive representations of sufficient coalitions

The explicit representation of sufficient coalitions can either use a boolean flag for each subset of N, or, more economically, only mention an antichain

[c] In this representation, a *dictatorship* is characterized by having a set of minimal winning coalition reduced to a singleton.

of minimally sufficient subsets. Whatever the chosen representation, it is not *compact*, as it potentially demands to store an amount of information that is exponential in the number of points of view.

Majority rule. A simplifying assumption consists in representing sufficient coalitions in an analogy to a voting setting: each criterion $i \in N$ is assigned a *voting power* $w_i \geq 0$ so that a given coalition of criteria $B \subseteq N$ is deemed sufficient if, and only if, its combined voting power $\sum_{i \in B} w_i$ is greater than a given *qualification threshold* λ.

$$\exists \lambda, \langle w_i \rangle_{i \in N} \in [0, +\infty[: \ \forall B \subseteq N, B \in \mathcal{T} \iff \sum_{i \in N} w_i \geq \lambda. \tag{4.9}$$

With this rule, the sufficient coalitions are represented in a compact form which is more amenable to linear programming.

The majority rule is used to represent sufficient coalitions both in Electre-Tri [Roy, 1991, Greco et al., 2010a] and most variants of the MR-Sort model [Leroy et al., 2011]. As a remarkable exception, [Sobrie et al., 2015] considers an extension of the majority rule where the voting power of the points of view are replaced by a *capacity*, a function mapping coalitions of points of view to nonnegative real numbers that is nondecreasing w.r.t. set inclusion. A coalition is considered sufficient if, and only if, its capacity exceeds the qualification threshold. This variant of the MR-Sort model is obviously equivalent to the NCS model, it offers two additional features:

- a numeric representation for the NCS model, that remains linear and can be handled by techniques and pieces of software dedicated to linear algebra, such as the powerful mixed integer programming (MIP) solvers;
- the notion of using a general capacity can easily be restricted to using limited forms, called k-capacities that restrict interactions between points of view. As 1-capacities boil down to the additive form of the majority rule, the normative assumption of representing upsets with k-capacities is often called k-*additivity*. This nesting of assumptions, with $k = 1$ corresponding to the majority rule and $k = |N|$ corresponding to the general NCS model, goes along a progressive expansion of the numeric parameter space, as going from $(k-1)$ to k additivity requires k among $|N|$ additional parameters.

4.3.4 Description in other paradigms.

In [Bouyssou and Marchant, 2007b], the authors connect some aggregation models for sorting described in other places to a general form of 'decomposable

threshold models' [Goldstein, 1991], which is naturally and concisely described in the *aggregate then compare* paradigm, with any value function V compatible with dominance (i.e. non decreasing w.r.t. to every coordinate). They show this broad model can be equivalently described quite simply in the *compare then aggregate* paradigm, in the form of a 'relational model with nested relations and unique profile', or in the *rule-based* paradigm in the form of a 'pessimistic at-least decision rules model' [Greco et al., 2001a]. The noncompensatory sorting model NCS particularizes these models. In particular, it can equivalently be described in the *aggregate then compare* paradigm by using a value function in the form of a Sugeno integral [Marichal, 2000, Bouyssou and Marchant, 2007b, Brabant et al., 2018][d]. Moreover, part of the appeal of the NCS model lies in the narrative structure it imposes to the generic rule-based model.

4.3.5 Veto.

We note the NCS model is born in part from an aspiration of offering a clean theoretical base to the Electre-Tri method [Roy, 1991, Greco et al., 2010a]. Indeed, when the notion of *veto*—values that are so bad, according to some point of view, that they place an upper cap to the assigned category—is discarded, Electre-Tri appears as a particular case of the NCS model with unique set of sufficient coalitions and an additive representation of coalitions. Conversely, a natural extension of the noncompensatory model is to allow for the notion of veto. Such an extension is considered in [Bouyssou and Marchant, 2007b], which details an axiomatic characterization of this model.

[d] This fact has been argued twice [Slowinski et al., 2002, Bouyssou and Marchant, 2007b], and proved once.

5

SAT-based formulations for Inv-NCS

5.1 Introduction

This chapter exposes our formal contributions to the aggregation of noncompensatory preferences into a sorting of the alternatives in ordered, predefined categories. In the framework of *noncompensatory sorting*, established by [Bouyssou and Marchant, 2007a, Bouyssou and Marchant, 2007b], it contributes to the elicitation of the model—i.e. finding adequate parameters of the model fitting the given *preference information*—by tackling this problem from a yet unexplored direction: formulating the search for parameters in the language of propositional logic, in a form that can be fed to one of the powerful SAT solvers that have emerged during the last decade.

5.1.1 NCS and Inv-NCS

In this section, we define the *inverse noncompensatory sorting* problem Inv-NCS 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, that gives a positive answer if, and only if, there is a preference parameter of the noncompensatory 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.

Definition 5.1 (Instances of Inv-NCS)**.** *An* instance *of the Inv-NCS problem is a sextuple* $(N, \mathbb{X}, \langle \succsim_i \rangle_{i\in N}, \mathbb{X}^\star, \{C^1 \prec \cdots \prec C^p\}, \alpha)$ *where:*

- N *is a finite set of* points of view*;*
- $\mathbb{X}$ *is a set of* alternatives*;*
- $\langle \succsim_i \rangle_{i\in N} \in \mathbb{X}^2$ *are* preferences*—for each point of view* $i \in N$, $\succsim_i \subset \mathbb{X}^2$ *is a total pre-ordering of the alternatives according to this point of view;*
- $\mathbb{X}^\star \subset \mathbb{X}$ *is a finite set of* reference alternatives *;*
- $\{C^1 \prec \cdots \prec C^p\}$ *is a finite set of* categories *ordered by* level of requirement*; and*
- $\alpha : \mathbb{X}^\star \to \{C^1 \prec \cdots \prec C^p\}$ *is an* assignment *of the reference alternatives to the categories.*

When referring to such an instance, we shall often shorten this sextuple as 'α': points of view, alternatives and preferences are usually elements specified by the context and can safely be left in the background, without any risk of confusion; reference alternatives and categories are implicitly defined, respectively, by the domain and codomain of the assignment α.

Definition 5.2 (Parameters of NCS)**.** *Given a context, a* parameter ω *of the NCS model is a couple* $(\langle \mathcal{A}_i^k \rangle_{i\in N,\ k\in[2..p]}, \langle \mathcal{T}^k \rangle_{k\in[2..p]})$*, where the* sufficient coalitions *satisfy* (4.1)

$$\forall k \in [2..p],\ \mathcal{T}^k \text{ is an upset of } (2^N, \subseteq) \text{ and } \mathcal{T}^2 \supseteq \cdots \supseteq \mathcal{T}^p;$$

and the approved sets *satisfy* (4.2)

$$\forall i \in N,\ \forall k \in [2..p],\ \mathcal{A}_i^k \text{ is an upset of } (\mathbb{X}_i, \precsim_i) \text{ and } \mathcal{A}_i^2 \supseteq \cdots \supseteq \mathcal{A}_i^p.$$

Sorting rule. Given a context and a parameter $\omega = (\langle \mathcal{A}_i^k \rangle_{i\in N,\ k\in[2..p]}, \langle \mathcal{T}^k \rangle_{k\in[2..p]})$, augmented with trivial values

$$\mathcal{T}^1 := \mathcal{P}(N),\ T^{p+1} := \varnothing, \quad \forall i \in N,\ \mathcal{A}_i^2 = \mathbb{X},\ \mathcal{A}_i^{p+1} = \varnothing,$$

NCS_ω is the function from $\mathbb{X}$ to $\{C^1 \prec \cdots \prec C^p\}$ satisfying (4.5):

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

Definition 5.3 (Solutions of Inv-NCS)**.** *Given a context, a* solution *of the instance* α *of the Inv-NCS problem is a parameter* ω *of the NCS model such that*

$$\forall x \in \mathbb{X}^\star,\ \alpha(x) = NCS_\omega(x). \tag{5.1}$$

5.1.2 Encoder and decoder.

A *Boolean function* is a function of the form $\{0,1\}^k \rightarrow \{0,1\}$ for a given nonnegative integer k. Any boolean function can be written in *conjunctive normal form*, as a conjunction of clauses where each clause is a disjunction of some variables or their negation. Such a function is said to be *satisfiable* if, and only if, 1 has at least one antecedent.

Precisely, this aim of this chapter is, given a context and an instance α of the Inv-NCS problem,to provide:

- an *encoding procedure* that yields a Boolean function Φ_α^{SAT} that is satisfiable iff there is a solution to the instance α of the Inv-NCS problem; and
- a *decoding procedure* that maps any antecedent of 1 by Φ_α^{SAT} to a solution ω of the instance α of the Inv-NCS problem.

5.1.3 A tale of two formulations.

The two contributions are based on Boolean satisfiability formulations:

- Section 5.2 introduces and extends a formulation originally described in [Belahcene et al., 2018c]. It is based on an explicit representation of the parameter space of the NCS model—coalitions of points of view and approved sets of alternatives, for each point of view and each level of requirement—leading to a formulation in conjunctive normal form with $O(2^{|N|} + p \times |N| \times |\mathbb{X}^\star|)$ variables and $O(p \times |\mathbb{X}^\star| \times 2^{|N|})$ clauses;
- Section 5.3 introduces and extends a formulation originally described in [Belahcene et al., 2018a]. 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 $O(p \times |N| \times |\mathbb{X}^\star|^2)$ variables and clauses;
- Section 5.4 addresses the issue of the *computational complexity* of the Inv-NCS problem, and concludes to its NP-hardness.

5.2 A SAT formulation for Inv-NCS based on coalitions

This section describes and extends a SAT formulation for Inv-NCS initially given in [Belahcene et al., 2018c].

5.2.1 Informal presentation of the approach

The formulation Φ_α^{SAT-C} yielded by the encoding presented in this section is based on an explicit representation of the parameter space of the noncompensatory sorting model—the pairs composed of a sequence of approved sets and a sequence of sufficient coalitions.

Variables. The Boolean function Φ_α^{SAT-C} operates on two types of variables:

- the 'a' variables, indexed by a point of view $i \in N$, a level of requirement $k \in [2..p]$ and a reference value $x \in \mathbb{X}^\star$, represent the approved sets $\mathcal{A}$, with the following semantic:

$$a_{i,k,x} = 1 \iff x \in \mathcal{A}_i^k \quad \text{i.e. } x \text{ is approved at level } k \text{ according to } i; \tag{5.2}$$

- the 't' variables, indexed by a coalition of points of view $B \subset N$ and a level of requirement $k \in [2..p]$, represent the sufficient coalitions $\mathcal{T}$, with the following semantic:

$$t_{B,k} = 1 \iff B \in \mathcal{T}^k \quad \text{i.e. the coalition } B \text{ is sufficient at level } k; \tag{5.3}$$

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

- Clauses ϕ_α^{C1} ensure each approved set $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}^\star, \precsim_i)$: if for a point of view i and a level of requirement k, the value x is approved, then any value $x' \succsim_i x$ must also be approved.
- Clauses ϕ_α^{C2} ensure approved sets are ordered by set inclusion according to their level of requirement: if an alternative x is approved at level k according to the point of view i, it should also be approved at level $k' < k$.
- Clauses ϕ_α^{C3} ensure each set of sufficient coalitions $\mathcal{T}$ is an upset for inclusion: if a coalition B is deemed sufficient at the level of requirement k, then a stronger coalition $B' \supset B$ should also be deemed sufficient at this level.
- Clauses ϕ_α^{C4} ensure sets of sufficient coalitions are ordered by inclusion according to their leve of requirementl: if a coalition B is deemed insufficient at level k, it should also be at any level $k' > k$.
- Clauses ϕ_α^{C5} ensure each alternative is not approved by a sufficient coalition of criteria at a level of requirement above the one corresponding to its assigned category.

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

Results Section 5.2.2 is devoted to the formal definition of the formulation Φ_α^{SAT-C} just described. While the role of clauses $\phi_\alpha^{C1}, \phi_\alpha^{C2}, \phi_\alpha^{C3}$ and ϕ_α^{C4} is straightforward with respect to the constraints defining ω is an acceptable parameter for the NCS model, the fact that the fulfillment of clauses ϕ_α^{C5} and ϕ_α^{C6} is both necessary and sufficient to ensure that $NCS_\omega \equiv \alpha$ is less so, and deserves close inspection in Section 5.2.3. Then Section 5.2.4 formalizes a decoding procedure permitting to leverage a solution of the SAT problem Φ_α^{SAT-C} into providing the parameter ω sought for eliciting the NCS model.

5.2.2 A SAT encoding of an instance of Inv-NCS

Definition 5.4. *Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \cdots \prec C^p\}$, we define the boolean function $\Phi_\alpha^{SAT\text{-}C}$ with variables $\langle a_{i,k,x}\rangle_{i\in N,\ k\in[2..p],\ x\in\mathbb{X}^\star}$ and $\langle t_{B,k}\rangle_{B\subseteq N,\ k\in[2..p]}$, as the conjunction of clauses:*

$$\Phi_\alpha^{SAT\text{-}C} := \phi_\alpha^{C1} \wedge \phi_\alpha^{C2} \wedge \phi_\alpha^{C3} \wedge \phi_\alpha^{C4} \wedge \phi_\alpha^{C5} \tag{5.4a}$$

$$\phi_\alpha^{C1} := \bigwedge_{i\in N,\ k\in[2..p]} \bigwedge_{x' \succsim_i x\, \in\mathbb{X}^\star} (a_{i,k,x'} \vee \neg a_{i,k,x}) \tag{5.4b}$$

$$\phi_\alpha^{C2} := \bigwedge_{i\in N,\ k<k'\in[2..p],\ x\in\mathbb{X}^\star} (a_{i,k,x} \vee \neg a_{i,k',x}) \tag{5.4c}$$

$$\phi_\alpha^{C3} := \bigwedge_{B\,\subset B'\,\subseteq\, N,\ k\in[2..p]} (t_{B',k} \vee \neg t_{B,k}) \tag{5.4d}$$

$$\phi_\alpha^{C4} := \bigwedge_{B\subseteq N,\ k<k'\in[2..p]} (t_{B,k} \vee \neg t_{B,k'}) \tag{5.4e}$$

$$\phi_\alpha^{C5} := \bigwedge_{B\subseteq N,\ k\in[2..p]} \bigwedge_{x\in\alpha^{-1}(C^{k-1})} (\bigvee_{i\in B} \neg a_{i,k,x} \vee \neg t_{B,k}) \tag{5.4f}$$

$$\phi_\alpha^{C6} := \bigwedge_{B\subseteq N,\ k\in[2..p]} \bigwedge_{x\in\alpha^{-1}(C^{k})} (\bigvee_{i\in B} a_{i,k,x} \vee t_{N\setminus B,k}) \tag{5.4g}$$

Efficiently enforcing monotonicity. Clauses $\phi_\alpha^{C1}, \phi_\alpha^{C2}, \phi_\alpha^{C3}$ and ϕ_α^{C4} enforce some kind of monotonicity condition. As written, they are highly redundant, threatening computational efficiency. Indeed, it is sufficient to consider clauses where ordered elements are adjacent to each other:

- for clauses ϕ_α^{C1}, solely consider clauses where either x' is a successor of x in the partial order $\succ_i$, or alternatives indifferent to each other, i.e. $x' \sim_i x$;
- for clauses ϕ_α^{C2} and ϕ_α^{C4}, solely consider clauses where levels of requirement are consecutive, i.e. $k' = k + 1$;

- for clauses ϕ_α^{C1}, solely consider clauses where coalitions B' and B differ on only a single point of view, i.e. $B' \setminus B = \{i\}$.

Note that the formulation put to the test in Chapter 6 has been streamlined according to these rules.

Model variants. As discussed in Section 4.3, the NCS model has many variants. Φ_α^{SAT-C} can easily be modified to account for two popular restrictions of the model:

- unique profiles—drop the index k concerning the level of requirement for the 'a' variables, ignore the conjunction over levels of requirement for clauses ϕ_α^{C1}, and ignore clauses ϕ_α^{C2} altogether;
- unique set of sufficient coalitions—drop the index k concerning the level of requirement for the 't' variables, ignore the conjunction over levels of requirement for clauses ϕ_α^{C3} and ignore clauses, ϕ_α^{C4} altogether.

The original presentation of this formulation, in [Belahcene et al., 2018c], was aimed at advancing the state of the art concerning the elicitation of MR-Sort models (see Section 4.3.3), and therefore focuses on the *unique set of sufficient coalitions* variant. The experimental study exposed in Chapter 6 remains focused on this variant, which is more reasonable than the unrestricted case in terms of computation time and preference information needed to pinpoint the parameters.

5.2.3 From a solution of Inv-NCS to a solution of the SAT formulation

The aim of this section is to establish Φ^{SAT-C} as an onto encoder for the Inv-SAT problem: if an instance α is positive (i.e. has a solution), then Φ_α^{SAT-C} should be satisfiable.

Theorem 5.1. *Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \cdots \prec C^p\}$, if the parameter $\omega = \big(\langle\ \mathcal{A}_i^k\ \rangle_{i\in N,\ k\in[2..p]},\ \langle\ \mathcal{T}^k\ \rangle_{k\in[2..p]}\big)$ is a solution of this instance, then the pair of tuples of binary values:*

- $\langle a_{i,k,x}\rangle_{i\in N,\ k\in[2..p],\ x\in\mathbb{X}^\star}$ *defined according to* (5.2)*; and*
- $\langle t_B\rangle_{B\subseteq N,\ k\in[2..p]}$ *defined by* (5.3)*;*

is mapped to 1 by the Boolean function Φ_α^{SAT-C}.

Proof. We check the satisfaction of every clause:

- The monotonicity clauses $\phi_\alpha^{C1}, \phi_\alpha^{C2}, \phi_\alpha^{C3}$ and ϕ_α^{C4} are satisfied by virtue of ω being a NCS parameter, so that the approved sets satisfy (4.2) and the sufficient coalitions satisfy (4.1). See the informal description of these types of clause in Section 5.2.1.

- For any exigence level $k \in [2..p]$, let $B \subseteq N$ a coalition of points of view and x an alternative assigned to C^{k-1} by α. Either $B \in \mathcal{T}^k$, and, by the sorting rule, $\{i \in N : x \in \mathcal{A}_i^k\} \not\subseteq B$, so the disjunction $\bigvee_{i \in B} \neg a_{i,k,x}$ is satisfied; or $B \notin \mathcal{T}^k$, and the atom $\neg t_{B,k}$ is satisfied. This proves the satisfaction of ϕ_α^{C5}.
- For any exigence level $k \in [2..p]$, any coalition $B \subset N$ and any alternative $x \in \mathbb{X}^\star$, suppose the disjunction $(\bigvee_{i \in B} a_{i,k,x} \vee t_{N \setminus B,k})$ is not satisfied. Hence, at exigence level k, x is disapproved according to all points of view in B, so the coalition approving x is at most as strong as $N \setminus B$, which is insufficient. Therefore, x does not meet the conditions to be assigned to category C^k or better by NCS. In particular, $x \notin \alpha^{-1}(C^k)$, which proves that ϕ_α^{C6} is satisfied.

□

Corollary 5.2. *Given an instance of Inv-NCS with an assignment α, if $\phi_\alpha^{SAT\text{-}C}$ is unsatisfiable, then α cannot be represented in the non compensatory sorting model.*

5.2.4 Decoding a solution of the SAT formulation into NCS parameters

The aim of this section is to provide a decoder for Φ_α^{SAT-C}, mapping any solution of the SAT formulation to a solution of Inv-NCS—i.e a parameter ω of the noncompensatory model such that the sorting functions NCS_ω and α coincide on the set of reference alternatives $\mathbb{X}^\star$.

Theorem 5.3. *Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \cdots \prec C^p\}$, if the binary tuple $\langle a_{i,k,x} \rangle_{i \in N,\ k \in [2..p],\ x \in \mathbb{X}^\star}$, $\langle t_B \rangle_{B \subseteq N,\ k \in [2..p]}$ is mapped to 1 by Φ_α^{SAT-C}, then the pair $\omega = (\langle\ \mathcal{A}_i^k\ \rangle_{i \in N,\ k \in [2..p]}, \langle\ \mathcal{T}^k\ \rangle_{k \in [2..p]})$ where:*

- *the approved sets $\langle\ \mathcal{A}_i^k\ \rangle_{i \in N,\ k \in [2..p]}$ are defined according to* (5.2)*; and*
- *the sufficient coalitions $\langle\ \mathcal{T}^k\ \rangle_{k \in [2..p]}$ are defined according to* (5.3)*;*

is a parameter of the noncompensatory sorting model. Moreover, ω is a solution of this instance.

Proof. Clauses ϕ_α^{C1} ensure each approved sets $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \precsim_i)$. Clauses ϕ_α^{C2} ensure approved sets are nested according to their level of requirement. Clauses ϕ_α^{C3} ensure each set of approved coalitions is an upset of $(\mathcal{P}(N), \subseteq)$. Clauses ϕ_α^{C4} ensure sets of sufficient coalitions are nested according to their level of requirement. Therefore, ω is a parameter for NCS.

Suppose ω is not a solution of the instance α of Inv-NCS. There is, at least, an alternative $x \in \mathbb{X}^\star$ such that $\alpha(x) \neq NCS_\omega(x)$.

- Suppose $\alpha(x) \succ NCS_\omega(x)$:
  x does not meet the demand of the sorting rule NCS_ω for the level of requirement $k : \alpha(x) = C^k$, thus $\{i \in N : x \in A_i^k\} \notin \mathcal{T}^k$. This violates the clause of ϕ_α^{C6} indexed by x, k and $B := \{i \in N : x \notin A_i^k\}$. Indeed, for all points of view $i \in B$, $a_{i,k,x} = 0$ by (5.2), and $N \backslash B = \{i \in N : x \notin A_i^k\} \notin T^k$, thus $t_{N\backslash B,k} = 0$ by (5.3).
- Suppose $\alpha(x) \prec NCS_\omega(x)$:
  x meets the demand of the sorting rule NCS_ω for the level of requirement $k : \alpha(x) = C^{k-1}$. Thus, $\{i \in N : x \in A_i^k\} \in \mathcal{T}^k$. This violates the clause of ϕ_α^{C5} indexed by x, k and $B := \{i \in N : x \in A_i^k\}$. Indeed, for all points of view $i \in B$, $a_{i,k,x} = 1$ by (5.2), and $t_{B,k} = 1$ by (5.3).

Either case would violate a clause of Φ_α^{SAT-C}. Therefore, ω is a solution of the instance α of Inv-NCS.

□

Corollary 5.4. *Given a context, the assignment α can be represented in the noncompensatory sorting model if, and only if, $\phi_\alpha^{SAT\text{-}C}$ is satisfiable.*

5.3 A SAT formulation based on pairwise separation conditions

The Boolean satisfiability formulation we propose in this section, denoted Φ_α^{SAT-P}, was originally described in [Belahcene et al., 2018a], with a focus on the case with two categories $C^1 \equiv$ Bad $\prec C^2 \equiv$ Good. We extend this formulation to the general case, with any number of categories, in Section 5.3.4. It is based on the following observations:

- The noncompensatory sorting model may appear particularly unwieldy to use explicitly, as it requires to handle explicitly the sufficient coalitions, which number is exponential in the number of points of view.
- In the case of the MR-Sort model, where coalitions are represented with the majority rule, in the case where the limiting profiles—and thus, the approved sets—are known, the problem of finding suitable parameters for the majority rule—the voting power of each point of view, and the majority threshold—boils down to a mere linear program (with continuous variables). This simple fact is leveraged in a heuristic approach to the inverse MR-Sort problem described in [Sobrie et al., 2015]: a limiting profile is guessed, 'best' fitting voting powers and majority threshold

are determined through LP minimization of a loss function, and the limiting profile is adjusted as to reach out to reference assignments not yet captured by the current parameters.

In the light of these observations, we begin, in Section 5.3.1 by focusing on the case where the approved sets are known. From this simple case, we derive a powerful characterization of the assignments that can be represented in the noncompensatory sorting model, detailed in Section 5.3.2. Section 5.3.3 proposes a SAT formulation in conjunctive normal form based on this characterization. Section 5.3.4 considers the extension of these results to cases with more than two categories.

5.3.1 Inv-NCS with fixed approved sets

When the approved sets are given, solving the inverse NCS problem – *i.e.* learning a set of sufficient coalitions permitting to represent the assignment in the noncompensatory sorting model – is similar to learning a disjunctive normal form from training examples. From this observation, we derive a polynomial time[a] algorithm yielding the *version space* [Mitchell, 1982] of the noncompensatory sorting model with fixed approved sets:

In this section, we consider given a context, an assignment $\alpha : \mathbb{X}^\star \to \{\textsc{Good}, \textsc{Bad}\}$ and a tuple $\langle \mathcal{A}_i \rangle_{i \in N}$ such that $\forall i \in N$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}_i, \succsim_i)$.

We define the following sets of coalitions:

$$\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^{\supseteq}_{\mathcal{P}(N)}\left(\bigcup_{g \in \alpha^{-1}(\textsc{Good})} \left\{\{i \in N : g \in \mathcal{A}_i\}\right\}\right), \tag{5.5}$$

$$\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^{\subseteq}_{\mathcal{P}(N)}\left(\bigcup_{b \in \alpha^{-1}(\textsc{Bad})} \left\{\{i \in N : b \in \mathcal{A}_i\}\right\}\right). \tag{5.6}$$

The following lemma explicit their roles with respect to Inv-NCS:

Lemma 5.5. *Given an instance of the Inv-NCS problem with an assignment $\alpha : \mathbb{X}^\star \to \{\textsc{Good}, \textsc{Bad}\}$, a tuple $\langle \mathcal{A}_i \rangle_{i \in N}$ such that $\forall i \in N$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}_i, \succsim_i)$, and an upset $\mathcal{S}$ of $(\mathcal{P}(N), \subseteq)$, the parameter $(\langle \mathcal{A}_i \rangle, \mathcal{S})$ is a solution of this instance if, and only if:*

$$\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \subseteq \mathcal{S} \subseteq \mathcal{P}(N) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) \tag{5.7}$$

Proof. α is represented by $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ iff i) for all alternatives $g \in \alpha^{-1}(\textsc{Good})$, $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(g) = \textsc{Good}$; and ii) for all alternatives $b \in \alpha^{-1}(\textsc{Bad})$, $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(b) = \textsc{Bad}$

[a] See Section 5.4.

i) holds iff $\mathcal{S}$ contains $\bigcup_{g\in\alpha^{-1}(\text{Good})}\{i \in N : g \in \mathcal{A}_i\}$ and, as a consequence of being an upset for inclusion, $\mathcal{S}$ contains the set $\mathcal{T}_{\langle\mathcal{A}_i\rangle}(\alpha)$ of all the coalitions at least as strong as those in $\bigcup_{g\in\alpha^{-1}(\text{Good})}\{i \in N : g \in \mathcal{A}_i\}$. ii) holds iff $\mathcal{S}$ does not contain any coalition pertaining neither to $\bigcup_{b\in\alpha^{-1}(\text{Bad})}\{i \in N : b \in \mathcal{A}_i\}$, nor to the set $\mathcal{F}_{\langle\mathcal{A}_i\rangle}(\alpha)$ of all the coalitions at least as weak as $\bigcup_{b\in\alpha^{-1}(\text{Bad})}\{i \in N : b \in \mathcal{A}_i\}$. □

This simple lemma is a very powerful tool for resolving instances of the Inv-NCS problem. It allows for features similar to the one permitted by the assumption of an additive representation of coalitions via the majority rule:

- It offers a simple representation of the *version space* of the $\mathcal{T}$ parameter—the set of sufficient coalitions—given the $\langle\mathcal{A}_i\rangle_{i\in N}$ parameter—the approved sets according to each point of view. The lower bound $\mathcal{T}_{\langle\mathcal{A}_i\rangle}(\alpha)$ is the set of the *necessarily sufficient coalitions* and $\mathcal{F}_{\langle\mathcal{A}_i\rangle}(\alpha)$ is the set of *necessarily insufficient coalitions*. Any coalition neither in one or the other is ambivalent: possibly sufficient and possibly insufficient. With the majority rule, the version space is a polytope in the parameter space.

- It allows to circumvent the encoding of the parameter $\mathcal{T}$ (the set of sufficient coalitions), which is a limiting factor of the efficiency of the formulation Φ_α^{SAT-C}. Instead, it opens the way for an approach consisting in encoding solely the $\langle\mathcal{A}_i\rangle_{i\in N}$ parameter (the approved sets), and find conditions that constrains the version space to be nonempty.

- Conversely, the lemma empowers the decoding of a partial solution $\langle\mathcal{A}_i\rangle_{i\in N}$ of an instance of the Inv-NCS problem into a suitable full parameter $(\langle\mathcal{A}_i\rangle_{i\in N}, \mathcal{T})$, with e.g. $\mathcal{T} := \mathcal{T}_{\langle\mathcal{A}_i\rangle}(\alpha)$ or $\mathcal{T} := \mathcal{P}(N)\setminus\mathcal{F}_{\langle\mathcal{A}_i\rangle}(\alpha)$, as both bounds are upsets of $(\mathcal{P}(N), \subseteq)$.

- Concerning accountability, it is noteworthy that unsuitable sets of sufficient coalitions—i.e upsets S of $(\mathcal{P}(N), \subseteq)$ that do not satisfy (5.7)—can be discarded in the light of *direct evidence*. Suppose S does not contain $\mathcal{T}_{\langle\mathcal{A}_i\rangle}(\alpha)$ (resp. is not disjoint to $\mathcal{F}_{\langle\mathcal{A}_i\rangle}(\alpha)$). There is a *witness* alternative g assigned to Good (resp. b assigned to Bad) that is approved by a coalition of points of view deemed insufficient (resp. sufficient), resulting in an obvious contradiction. Moreover—and this is not the case with the majority rule—an empty version space always results from a conflict between a Good and a Bad witness—a fact the theorem from the next section capitalizes on.

5.3.2 A pairwise characterization of positive instances

The following theorem implements the strategy sketched in the previous section and characterizes YES instances of the Inv-NCS problem with conditions tying solely the 'approved sets' component of the parameter space.

Theorem 5.6. *An assignment α of alternatives to categories can be represented in the noncompensatory sorting model if, and only if, there is a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^N$ such that:*

1. *for each point of view $i \in N$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$*
2. *for each pair of alternatives $(g, b) \in \alpha^{-1}(\text{Good}) \times \alpha^{-1}(\text{Bad})$, there is at least one point of view $i \in N$ such that $g \in \mathcal{A}_i$ and $b \notin \mathcal{A}_i$.*

This result is very important as it says that, in order to check that an assignment α is compatible with NCS, it is equivalent to find approval subsets over each point of view such that one can discriminate each pair of Good and Bad alternatives on at least one point of view (*i.e.* the Good alternative is approved on this point of view, and not the Bad one). Interestingly, the concept of sufficient coalitions disappears in this characterization.

Proof.

[$\neg$(1+2) $\Rightarrow$ $\neg$NCS] If there are two alternatives $g \in \alpha^{-1}(\text{Good})$ and $b \in \alpha^{-1}(\text{Bad})$ that falsify Condition 2, then, for any potential parameter $\omega = (\langle \mathcal{A}_i \rangle_{i \in N}, \mathcal{S})$ of a noncompensatory sorting model, the nesting $\{i \in N : g \in \mathcal{A}_i\} \subseteq \{i \in N : b \in \mathcal{A}_i\}$ results in a sorting NCS_ω at least as favorable to b as to g, whereas $\alpha(b) = \text{Bad}$ is strictly worse than $\alpha(g) = \text{Good}$.

[(1+2) $\Rightarrow$ NCS] Given a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^N$ satisfying conditions 1 and 2, we consider the sets of coalitions $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$.

According to Lemma 5.5, α can be represented in the noncompensatory model iff $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) = \varnothing$. Suppose this intersection is nonempty, and let $B \in \mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$. By definition of $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$, there is an alternative $g \in \alpha^{-1}(\text{Good})$ such that $B \supseteq \{i \in N : g \in \mathcal{A}_i\}$: for all points of view $i \notin B$, $g \notin \mathcal{A}_i$. By definition of $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$, there is an alternative $b \in \alpha^{-1}(\text{Bad})$ such that $B \subseteq \{i \in N : b \in \mathcal{A}_i\}$: for all points of view $i \in B$, $b \in \mathcal{A}_i$. Consequently, there is no point of view according to which g is accepted but not b, contradicting condition 2. Hence, $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) = \varnothing$. □

5.3.3 A compact SAT formulation for Inv-NCS

We leverage Theorem 5.6 by formulating a Boolean satisfiability problem Φ_α^{SAT-P} that answers the decision problem: 'can the assignment α be represented in the noncompensatory model?'

Encoding. Similarly to the formulation Φ_α^{SAT-C} described is Section 5.2, the formulation Φ_α^{SAT-P} operates on two types of variables:

- 'a' variables, representing the approved sets, with the exact same semantics as their counterpart in Φ_α^{SAT-C}, i.e.

$$a_{i,x} = \begin{cases} 1 & \text{if } x \in \mathcal{A}_i \text{ i.e. } x \text{ is approved according to } i; \\ 0 & \text{else.} \end{cases}$$

- auxiliary 's' variables, indexed by a point of view $i \in N$, 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 the point of view i, i.e.

$$s_{i,g,b} = \begin{cases} 1 & \text{if } g \in \mathcal{A}_i \text{ and } b \notin \mathcal{A}_i; \\ 0 & \text{else.} \end{cases}$$

Φ_α^{SAT-P} is the conjunction of four types of clauses: ϕ_α^{P1} ensuring each $\mathcal{A}_i$ is an upset (matching the first condition of Theorem 5.6), 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 point of view (matching the second condition of Theorem 5.6).

Definition 5.5. *Given an instance of Inv-NCS with two categories and an assignment $\alpha : \mathbb{X}^\star \to \{\text{Bad} \prec \text{Good}\}$, we define the boolean function $\Phi_\alpha^{SAT\text{-}P}$ with variables $\langle a_{i,x} \rangle_{i \in N,\ x \in \mathbb{X}^\star}$ and $\langle s_{i,g,b} \rangle_{i \in N,\ g \in \alpha^{-1}(\text{Good}),\ b \in \alpha^{-1}(\text{Bad})}$, as the conjunction of clauses:*

$$\phi_\alpha^{SAT\text{-}P} := \phi_\alpha^{P1} \wedge \phi_\alpha^{P2} \wedge \phi_\alpha^{P3} \wedge \phi_\alpha^{P4} \tag{5.8}$$

$$\begin{aligned}
\phi_\alpha^{P1} &:= \bigwedge_{i \in N} \bigwedge_{x' \succsim_i x \in \mathbb{X}^\star} (a_{i,x'} \vee \neg a_{i,x}) && \text{(5.9a)} \\
\phi_\alpha^{P2} &:= \bigwedge_{i \in N,\ g \in \alpha^{-1}(\text{Good}),\ b \in \alpha^{-1}(\text{Bad})} (\neg s_{i,g,b} \vee \neg a_{i,b}) && \text{(5.9b)} \\
\phi_\alpha^{P3} &:= \bigwedge_{i \in N,\ g \in \alpha^{-1}(\text{Good}),\ b \in \alpha^{-1}(\text{Bad})} (\neg s_{i,g,b} \vee a_{i,g}) && \text{(5.9c)} \\
\phi_\alpha^{P4} &:= \bigwedge_{g \in \alpha^{-1}(\text{Good}),\ b \in \alpha^{-1}(\text{Bad})} (\bigvee_{i \in N} s_{i,g,b}) && \text{(5.9d)}
\end{aligned}$$

The formulation is compact: $O(|N|.|\mathbb{X}|^2)$ variables, $O(|N|.|\mathbb{X}|^2)$ binary clauses and $O(|\mathbb{X}|^2)$ $|N|$-ary clauses.

Decoding. If the instance is a YES, any solution of the satisfiability problem translates into suitable, yet arbitrary, explicit values for the approved sets. Upper and lower bounds for the set of sufficient coalitions can then be obtained thanks to Lemma 5.5.

Corollary 5.7. *Given an instance of Inv-NCS with two categories and an assignment $\alpha : \mathbb{X}^\star \to \{\text{BAD} \prec \text{GOOD}\}$, this instance is positive if, and only if, $\phi_\alpha^{SAT\text{-}P}$ is satisfiable.*

Moreover, if $\langle a_{i,x}\rangle, \langle s_{i,g,b}\rangle$ is an antecedent of 1 by $\phi_\alpha^{SAT\text{-}P}$, then the parameter $\omega := (\langle \mathcal{A}_i\rangle, \mathcal{S})$ with accepted sets defined by $\mathcal{A}_i := \{x \in \mathbb{X} : a_{i,x} = 1\}$ and any upset $\mathcal{S}$ of $(\mathcal{P}(N), \subseteq)$ of sufficient coalitions containing the upset $\mathcal{T}_{\langle \mathcal{A}_i\rangle}(\alpha)$ and disjoint from the lower set $\mathcal{F}_{\langle \mathcal{A}_i\rangle}(\alpha)$ is a solution of this instance.

5.3.4 More than two categories

The case where there are $p > 2$ categories $\{C^1 \prec \cdots \prec C^p\}$ requires a few adaptations of the formulation given in the preceding section, and presented in [Belahcene et al., 2018a]. It relies mostly on the fact that a NCS model with p categories is, informally, the combination of $p-1$ NCS models with two categories which parameters satisfy the nesting conditions (4.1) and (4.2).

Given an assignment α and a level of requirement $k \in [2..p]$, we define the set of alternatives assigned at least to C^k as

$$C^{\succeq k} := \bigcup_{k' \in [k..p]} C^{k'}. \tag{5.10}$$

We propose the following definition for $\Phi_\alpha^{SAT-P'}$, that coincides with Φ_α^{SAT-P} (see Definition 5.5) when $p = 2$.

Definition 5.6. *Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \cdots \prec C^p\}$, we define the boolean function $\Phi_\alpha^{SAT\text{-}P'}$ with variables $\langle a_{i,k,x}\rangle_{i\in N,\ k\in[2..p], x\in\mathbb{X}^\star}$ and $\langle s_{i,k,g,b}\rangle_{i\in N,\ k\in[2..p],\ g\in\alpha^{-1}(G^{\succeq k}),\ b\notin\alpha^{-1}(G^{\succeq k})}$, as the conjunction of clauses:*

$$\Phi_\alpha^{SAT\text{-}P'} := \phi_\alpha^{P'1} \wedge \phi_\alpha^{P'2} \wedge \phi_\alpha^{P'3} \wedge \phi_\alpha^{P'4} \wedge \phi_\alpha^{P'5} \tag{5.11}$$

$$\phi_\alpha^{P'1} := \bigwedge_{i\in N,\ k\in[2..p]} \bigwedge_{x' \succsim_i x \in \mathbb{X}^\star} (a_{i,k,x'} \vee \neg a_{i,k,x}) \tag{5.12a}$$

$$\phi_\alpha^{P'2} := \bigwedge_{i\in N,\ k<k'\in[2..p],\ x\in\mathbb{X}^\star} (a_{i,k,x} \vee \neg a_{i,k',x}) \tag{5.12b}$$

$$\phi_\alpha^{P'3} := \bigwedge_{i\in N,\ k\in[2..p]} \bigwedge_{g\in\alpha^{-1}(C^{\succeq k}),\ b\notin\alpha^{-1}(C^{\succeq k})} (\neg s_{i,k,g,b} \vee \neg a_{i,k,b}) \tag{5.12c}$$

$$\phi_\alpha^{P'4} := \bigwedge_{i\in N,\ k\in[2..p]} \bigwedge_{g\in\alpha^{-1}(C^{\succeq k}),\ b\notin\alpha^{-1}(C^{\succeq k})} (\neg s_{i,k,g,b} \vee a_{i,k,g}) \tag{5.12d}$$

$$\phi_\alpha^{P'5} := \bigwedge_{k\in[2..p]} \bigwedge_{g\in\alpha^{-1}(C^{\succeq k}),\ b\in\alpha^{-1}(C^{\succeq k})} (\bigvee_{i\in N} s_{i,k,g,b}) \tag{5.12e}$$

The remarks made in Section 5.2.2 about an efficient implementation of Φ_α^{SAT-C} apply here: many clauses are redundant in $\phi_\alpha^{P'1}$ and $\phi_\alpha^{P'2}$ and can safely be ignored.

Theorem 5.8. *Given a context, an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \cdots \prec C^p\}$ can be represented in the noncompensatory sorting model if, and only if, $\phi_\alpha^{SAT\text{-}P'}$ is satisfiable.*

Moreover, if $\langle a_{i,k,x}\rangle, \langle s_{i,k,g,b}\rangle$ is an antecedent of 1 by $\phi_\alpha^{SAT\text{-}P}$, then the parameter $\omega := (\langle \mathcal{A}_i^k\rangle, \langle \mathcal{T}^k\rangle)$ with accepted sets defined by $\mathcal{A}_i^k := \{x \in \mathbb{X} : a_{i,k,x} = 1\}$ and sufficient coalitions defined by

$$\mathcal{T}^k := cl^{\supseteq}_{\mathcal{P}(N)}\Big(\bigcup_{g \in \alpha^{-1}(C^{\geq k})} \{i \in N : g \in \mathcal{A}_i^k\}\Big) \tag{5.13}$$

is a solution of the instance α of the Inv-NCS problem.

Proof. The proof relies on the following lemma, which is merely a reformulation of the definition of the noncompensatory sorting model:

Lemma 5.9. *The instance $\alpha : \mathbb{X}^\star \to \{C^1 \prec \cdots \prec C^p\}$ is a positive instance of Inv-NCS if, and only if, all the instances $\langle \alpha_k : \mathbb{X}^\star \to \{\textsc{Bad} \prec \textsc{Good}\}\rangle_{k\in[2..p]}$ with two categories, defined by*

$$\alpha_k(x) = \begin{cases} \textsc{Good} & \text{if } \alpha(x) \in C^{\geq k}; \\ \textsc{Bad} & \text{else.} \end{cases} \tag{5.14}$$

are positive instances of Inv-NCS and have respective solutions $\omega_k = (\langle \mathcal{A}_i^k\rangle_{i\in N}, \mathcal{T}^k)$ that collectively satisfy the nesting conditions (4.1) *and* (4.2).

By construction, $\Phi_\alpha^{SAT-P'} = \bigwedge_{k\in[2..p]} \Phi_{\alpha_k}^{SAT-P} \wedge \phi_\alpha^{P'2}$, ensuring $\Phi_\alpha^{SAT-P'}$ is satisfiable if, and only if, the α_k are all positive instances of Inv-NCS, with nested satisfactory values (enforced by $\phi_\alpha^{P'2}$). Moreover, Lemma 5.5 guarantees that each set $\mathcal{T}^k$ defined by (5.13) yields, when combined with the corresponding tuple $\langle \mathcal{A}_i^k\rangle_{i\in N}$ of approved sets, a parameter solution of the instance α_k. The only verification left is whether the sets of sufficient coalitions $\langle \mathcal{T}^k\rangle_{k\in[2..p]}$ are effectively nested. This fact is established by the following arguments:

i) the upper closure is an isotonic operator;

ii) the union is taken on a set of alternatives $g \in \alpha^{-1}(C^{\geq k})$ that decreases as the level of requirement k increases; and

iii) for a given good alternative g, the sequence of sets $\langle \{i \in N : g \in \mathcal{A}_i^k\}\rangle_{k\in[2..p]}$ is decreasing because the sequence of sets $\langle \mathcal{A}_i^k\rangle_{k\in[2..p]}$ is decreasing.

□

5.4 Computational complexity of Inv-NCS

This section addresses the question of the intrinsic computational complexity of the problem Inv-NCS.

Boolean satisfiability offers a powerful language permitting to describe difficult combinatorial problems [Cook, 1971]. When written in conjunctive normal form, SAT instances can be given wholesale to dedicated solvers, that eschews the need for developing a dedicated piece of software and benefits from state-of-the-art refinements in the solving of such problems. Nevertheless, it would be unwise to delegate the search for a parameter of the NCS model consistent with a given assignment to such a solver, if this search were not, intrinsically, a difficult combinatorial problem. This section addresses this issue.

5.4.1 Complexity of Inv-NCS with fixed approved sets

We begin by a simple, yet strong result for the simplified version of Inv-NCS where the approved sets are given, and the question asked boils down to: *'Are there nested sets of sufficient coalitions such that a given assignment can be represented in the noncompensatory sorting model?'*

Corollary 5.10 (complexity of Inv-NCS with fixed approved sets)**.** *Given an instance of Inv-NCS with an assignment α of alternatives to categories and a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \precsim_i) \rangle$, the problem of deciding whether α can be represented in the noncompensatory sorting model with approved sets $\langle \mathcal{A}_i \rangle$ is polynomial time.*

Proof. By Lemma 5.5, it boils down to checking whether $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is empty or not, which is $O(|\mathbb{X}|^2.|N|)$. □

This result has far-reaching implications. Inv-NCS aims at retrieving a correct parameter for NCS. Because the parameters for NCS mention a set of sufficient coalitions, the representation of a parameter is potentially non-compact with respect to the instance α of the problem, whose size is linear in the number of points of view. Therefore, Inv-NCS does not belong to the class NP, as checking a potential solution with explicit sufficient coalitions requires to enumerate over the power set of the points of view. Nevertheless, the version space of Inv-NCS can be checked in polynomial time! This places the NCS model with given approved sets in the peculiar position where the version space of the model is simpler to store and access than a fully defined representative model, a fact we leverage in Chapter 7.

5.4.2 NP-hardness of Inv-NCS

We show that a solver fo the inverse NCS problem is able to solve any problem of the class NP. Assuming P ≠ NP, it means that Inv-NCS is intractable: there is no algorithm able to solve every instance in a number of steps bounded by a polynomial in the size of the instance.

Theorem 5.11. *The problem of deciding whether an instance α can be represented in the noncompensatory sorting model is NP-hard, even in the case where there are only two categories.*

Proof. By reduction from SAT: Given an instance of the SAT problem—written in conjunctive normal form, with n variables $y^1, \dots, y^n$ and m clauses $c_1 \wedge \dots \wedge c_m$—we build a gadget instance of Inv-NCS of a size bounded by a polynomial in the size of the SAT instance, such that solving the Inv-NCS problem with this instance permits to solve the original instance of the SAT problem. Therefore, any Inv-NCS solver can serve as a SAT solver, and the Inv-NCS problem is at least as hard as the SAT problem.

We build our gadget assignment with $m+n$ points of view $N := [1..(m+n)]$, $2m$ alternatives $\mathbb{X}^\star := \{g_1, \dots, g_m, b_1, \dots, g_m\}$ and two categories $\{\textsc{Bad} \prec \textsc{Good}\}$. $g_1, \dots, g_m$ are assigned to Good whereas $b_1, \dots, b_m$ are assigned to Bad.

First, let us focus on the first m points of view: for eack $j \in 1 \dots m$, let $g_j \sim_j b_j \succ_j g_1 \sim_j \dots \sim_j g_{j-1} \sim_j g_{j+1} \sim_j \dots \sim_j g_m \sim_j b_1 \sim_j \dots \sim_j b_{j-1} \sim_j b_{j+1} \sim_j \dots \sim_j b_m$. The preference $\succsim_j$ has two equivalence classes, the upper one containing $\{g_j, b_j\}$ and the lower one containing $\bigcup_{j' \neq j} \{g_{j'}, b_{j'}\}$. The n last points of view of the gadget are built considering the SAT formula.

The last n points of view are built according to the clauses. Each clause is a disjunction of atoms—either a variable y_j for some index $j \in [1..n]$ or its negation $\neg y_j$. For $j \in [1..n]$, define P_j as the subset of clauses containing the atom y_j, and N_j as the subset of clauses containing the atom $\neg y_j$. The preference relation $\succsim_{j+m}$ is constructed so as to have at most 3 equivalence classes: the uppermost containing the alternatives $\bigcup_{l \in P_j} \{g_l\}$, the one in the middle containing $\bigcup_{l \in P_j} \{b_l\} \cup \bigcup_{l \in N_j} \{g_l\}$, and the lowest containing $\bigcup_{l \in N_j} \{b_l\} \cup \bigcup_{l \notin P_j \cup N_j} \{g_l, b_l\}$.

We use the pairwise characterization of positive instances of Inv-NCS with two categories offered by Theorem 5.6. Furthermore, we note trivial accepted sets—i.e. points of view $i \in N$ such that $\mathcal{A}_i = \varnothing$ or $\mathcal{A}_i = \mathbb{X}^\star$—do not contribute to the feasibility of the inverse NCS problem.

For the n first points of view, there is only one nontrivial accepted set: it accepts the upper class and rejects the lower one. For the n last points of view of the gadget, the nontrivial accepted sets accept the uppermost equivalence

class, reject the lowest class, and either accept or reject the class in the middle. We define a one-to-one mapping between the nontrivial accepted sets of the gadget and the assignment of the n variables of the SAT problem:

$$y_j := \begin{cases} 1, & \text{if } A_{n+j} = \bigcup_{l \in P_j} \{g_l\} \\ 0, & \text{if } \mathcal{A}_{n+j} = \bigcup_{l \in P_j} \{g_l, b_l\} \cup \bigcup_{l \in N_j} \{g_l\}. \end{cases}$$

Thus, the assignment of truth values to the variables y_j defined in the remainder of the proof (applying the previous patch) ensures that :

[The pair (g_j, b_j) is discriminated according to p.o.v. k] if, and only if, [[[y_k is a positive atom appearing in c_j] and [y_k is assigned to one]] or [[$\neg y_k$ is a negative atom appearing in c_j] and [y_k is assigned to zero]]]

Therefore, the pair (g_j, b_j) is discriminated according to some p.o.v. if, and only if, the clause c_j is satisfied: a solution of the SAT problem is mapped to a tuple of accepted sets that discriminates all pairs with opposite assignments and reciprocally.

□

6

Experimental results concerning the resolution of Inv-NCS via SAT

6.1 Introduction

This chapter is devoted to the assessment of the SAT formulations proposed in Chapter 5.

6.1.1 Context

When trying to compare the novel SAT formulations established in Chapter 5 to the state of the art concerning the elicitation of noncompensatory sorting models, great care must be taken concerning the specific assumptions made about the model.

On the one hand, the novel formulations are geared towards the elicitation of *noncompensatory sorting* models (NCS) from learning examples, and can handle variants of NCS where the space of parameters are restricted (see Section 4.3.2).

On the other hand, the state of the art concerning the elicitation of NCS models is concerned with variants of the model, called *majority rule sorting* (MR-Sort) [Leroy et al., 2011], where:

- there is only one set of sufficient coalitions for all levels of requirement (see Section 4.3.2); and

- sufficient coalitions have an additive representation with the *majority rule* (see Section 4.3.3).

These restrictions stem from a combination of issues

- legacy, as to remain close to the Electre Tri foundational model [Roy, 1991, Greco et al., 2010a];
- representation, as to remain compatible with the mixed integer programming tools often used in Operations Research; and
- performance, as even with this straitjacket of restrictions made to the NCS model, only toy instances with very few points of views and learning examples can be solved in a reasonable amount of time.

6.1.2 Research question

The main question this experimental study is trying to answer is:

'To what extent can the SAT formulations for Inv-NCS can contribute to learning the parameter of a noncompensatory sorting model from actual assignment examples?'

The theoretical soundness of these formulations is not in question. It is addressed in Chapter 5, and know for a fact that these formulations are logically faithful to the elicitation problem. Nevertheless, the scope of the research question is quite broad.

- *Computational efficiency:* is the actual computation time of these formulations acceptable? How does it compare to the state of the art? How do the two formulations compare to each other?
- *Relevance:* How close to the *ground truth* is the corresponding model, compared to the one obtained with state-of-the-art methods? This question is actually twofold:
  - is it wise to do without any loss function?
    The SAT formulations are inherently geared to yield, if the problem is feasible, an arbitrary parameter for the NCS model, while the state-of-the-art formulations are based on an Optimization framework and return the 'most representative model' in the sense of some loss function.
  - is it wise to do without any model parsimony?
    The NCS class of sorting procedure elicited by the SAT formulations is much broader than the MR-Sort class. While this feature is a blessing in terms of expressiveness—it broaden the spectrum of decisional behaviors captured by the model—it comes with a cost for elicitation, both informational, as more learning examples are

likely needed to pinpoint the ground truth, and computational, as the size of the instances increases.

6.1.3 Layout of the chapter

The remainder of the chapter is organized in three section.

- Section 6.2 briefly recalls the state-of-the-art formulation, that we denote Φ^{MIP-O}, used to solve the inverse MR-Sort problem, based on mixed integer programming, and proposed by [Leroy et al., 2011]. We also propose a modified version, denoted Φ^{MIP-D}, obtained by turning off the optimization of the loss function, in order to assess its influence—both positive and negative—on the elicitation.
- Section 6.3 exposes the experimental protocol and the results of a study devoted to comparing the formulations Φ^{SAT-C}, Φ^{MIP-O} and Φ^{MIP-D}, originally presented in [Belahcene et al., 2018c].
- Section 6.4 discusses the results of this experimental studies, in the perspective of the research questions.

6.2 Learning MR-Sort using Mixed Integer Programming

Learning the parameters of an MR-Sort model using mixed integer programming has been studied in [Leroy et al., 2011]. We recall here the method used in [Leroy et al., 2011] in order to obtain the mixed integer program (MIP) formulation that infers an MR-Sort model on the basis of examples of assignments.

With MR-Sort (see Definition 3), the condition for an alternative $x \in \mathbb{X}^\star$ to be assigned to a category C^h reads:

$$x \in C^h \iff \begin{cases} \sum_{i=1}^{n} c_{x,i}^{h-1} & \geq \lambda \\ \sum_{i=1}^{n} c_{x,i}^{h} & < \lambda \end{cases} \quad \text{with} \quad c_{x,i}^{k} = \begin{cases} w_i & \text{if } x_i \geq b_i^k, \\ 0 & \text{otherwise.} \end{cases}$$

The linearization of these constraints induces the use of binary variables. For each variable $c_{x,i}^k$, with $k = \{h-1, h\}$, we introduce a binary variable $\delta_{x,i}^k$ that is equal to 1 when the performance of $x \in \mathbb{X}^\star$ is at least as good as or better than the performance of b^k on the criterion i and 0 otherwise. For an alternative x assigned to a category C^h with $2 \leq h \leq p-1$, it introduces $2n$ binary variables. For alternatives assigned to one of the extreme categories, the number of binary variables is divided by two. The value of each variable

$\delta^k_{x,i}$ is obtained thanks to the following constraints:

$$M(\delta^k_{x,i} - 1) \leq x_i - b^k_i < M \cdot \delta^k_{x,i} \tag{6.1a}$$

in which M is an sufficiently large positive constant[a]. The value of $c^k_{x,i}$ are finally obtained thanks to the following constraints:

$$\begin{cases} 0 \leq & c^k_{x,i} & \leq w_i, \\ \delta^k_{x,i} - 1 + w_i \leq & c^k_{x,i} & \leq \delta^k_{x,i}. \end{cases} \tag{6.1b}$$

The dominance structure on the set of profiles is ensured by the following constraints:

$$\forall i \in N, h = \{2, \ldots, p-1\},\ b^h_i \geq b^{h-1}_i \tag{6.1c}$$

As the equation (4.9) defining the majority rule is homogenous, the coefficients $\langle w \rangle$ and λ can be multiplied by any positive constant without modifying the upset of coalitions they represent. Thus, the following normalization constraint can be added without loss of generality:

$$\sum_{i=1}^{n} w_i = 1. \tag{6.1d}$$

To obtain a MIP formulation, the next step consists to define an objective function. In [Leroy et al., 2011], two objective functions are considered, one of which consists in maximizing the robustness of the assignments. It is done by adding continuous variables $\mu_x, \nu_x \in \mathbb{R}$ for each alternative $x \in \mathbb{X}^\star$ such that:

$$\begin{cases} \sum_{i=1}^{n} c^{h-1}_{x,i} & = \lambda + \mu_x, \\ \sum_{i=1}^{n} c^{h}_{x,i} & = \lambda - \nu_x. \end{cases} \tag{6.1e}$$

The objective function consists in optimizing a slack variable σ that is constrained by the values of the variables μ_x and ν_x as follows:

$$\forall x \in \mathbb{X}^\star, \begin{cases} \sigma & \leq \mu_x, \\ \sigma & \leq \nu_x. \end{cases} \tag{6.1f}$$

The combination of the objective function and all the constraints listed above leads to MIPs that can be found in [Leroy et al., 2011].

Definition 6.1 (MIP-O formulation for MR-Sort)**.** *Given an assignment $\alpha : \mathbb{X}^\star \to \{1 \prec \cdots \prec p\}$, we denote Φ^{MIP-O}_α the mixed linear program with decision variables σ, λ, $\langle b^k_i \rangle_{i \in N, k \in [1..p-1]}$, $\langle w_i \rangle_{i \in N}$, $\langle c^h_{x,i} \rangle_{i \in N, x \in \mathbb{X}^\star, h \in \{\alpha(x)-1, \alpha(x)\}}$, $\langle \mu_x \rangle_{x \in \mathbb{X}^\star}$, $\langle \nu_x \rangle_{x \in \mathbb{X}^\star} \in \mathbb{R}^+$ and $\langle \delta^h_{x,i} \rangle_{i \in N, x \in \mathbb{X}^\star, h \in \{\alpha(x)-1, \alpha(x)\}} \in \{0, 1\}$, consisting in minimizing the objective σ, subject to the constraints (6.1a), (6.1b), (6.1c), (6.1d), (6.1e) and (6.1f).*

[a] $M > Max_{i \in N} max \mathbb{X}_i$

Faithfulness of the MIP-O formulation

Proposition 6.1 ([Leroy et al., 2011]). *An assignment $\alpha : \mathbb{X}^\star \to \{1 \prec \cdots \prec p\}$ can be represented in the model MR-Sort if, and only if, Φ_α^{MIP-O} is feasible. If the tuple $\langle \sigma, \lambda, b, w, c, \mu, \nu, \delta \rangle$ is a feasible solution of Φ_α^{MIP-O}, then the tuple of profiles b, the tuple of voting powers w and the majority threshold λ are suitable parameters of a MR-Sort model that extends the assignment α.*

We are looking to compare this state-of-the-art formulation to the boolean satisfiability formulation Φ^{SAT-C} defined in Chapter 5 in terms of computational efficiency, and in terms of quality of the result. Yet, we suspect the two approaches differ in too many aspects to be meaningfully compared. The Φ^{MIP-O} formulation is based on a numerical representation of the problem, considers the set of every MR-Sort model extending the assignment, and selects the best according to the objective function – here, returning the model that gives the sharpest difference in voting weights between sufficient and insufficient coalitions of criteria. Meanwhile, the SAT formulations Φ^{SAT-C} are based on a logical representation of the problem, consider the wider set of every *noncompensatory sorting model with unique set of sufficient coalitions* extending the assignment, and randomly yields a suitable model. In order to be able to credit the effects we would observe to the correct causes, we introduce a third formulation, called Φ^{MIP-D}, that helps bridging the gap between Φ^{MIP-O} on one hand and Φ^{SAT-C} on the other hand. Φ^{MIP-D} is formally a mixed integer program with a null objective function. This trick enables us to use the optimization shell of the MIP formulations to express a decision problem assessing the satisfiability of the constraints, and yielding a random solution (which, in our context, represents a particular MR-Sort model), rather than looking for the best one in the sense of the objective function. Another instance of this configuration, where an optimization problem is compared to its feasibility version, can be found in [Dickerson et al., 2014].Here, it should be noted that the MIP-D formulation differ from the feasibility version of MIP-O on the way insufficient coalitions of criteria are characterized. Theoretically, insufficient coalitions are defined by a strict comparison, that cannot be represented directly in the linear optimization framework[b].

- The optimization version circumvents this obstacle by maximizing the contrast in normalized voting power between sufficient and insufficient coalitions. Finding a nonzero contrast guarantees that sufficient and insufficient coalitions can be strictly separated.

[b]Mathematical programming relies on the fact that the domain upon which the objective function is optimized is topologically closed. This would not be the case, should the comparisons implicitly defining this domain were allowed to be non-strict.

- The feasibility version addresses the obstacle by leaving the total weight unconstrained, but requires the minimal difference between sufficient and insufficient coalitions is at least one vote.

This slight difference might account for some divergence of behavior we observe during our experiment (see Section 6.3, and particularly 6.3.3).

Definition 6.2 (MIP-D formulation for MR-Sort)**.** *Given an assignment* $\alpha : \mathbb{X}^\star \to \{1 \prec \cdots \prec p\}$*, we denote* Φ_α^{MIP-D} *the mixed linear program with decision variables* $\langle b_i^k \rangle_{i\in N, k\in[1..p-1]}$, $\langle w_i \rangle_{i\in N}$, λ, $\langle \mu_x \rangle_{x\in\mathbb{X}^\star}$, $\langle \nu_x \rangle_{x\in\mathbb{X}^\star}$, $\langle c_{x,i}^h \rangle_{i\in N, x\in\mathbb{X}^\star, h\in\{\alpha(x)-1,\alpha(x)\}} \in \mathbb{R}^+$ *and* $\langle \delta_{a,i}^h \rangle_{i\in N, x\in\mathbb{X}^\star, h\in\{\alpha(x)-1,\alpha(x)\}} \in \{0,1\}$*, consisting in minimizing the objective 0, subject to the constraints (6.1a), (6.1b), (6.1c), (6.1e) and (6.1g), where:*

$$\forall x \in \mathbb{X}^\star, \begin{cases} 1 & \leq \mu_x, \\ 1 & \leq \nu_x. \end{cases} \tag{6.1g}$$

Theorem 6.2 (Faithfulness of the MIP-D formulation)**.** *An assignment* $\alpha : \mathbb{X}^\star \to \{1 \prec \cdots \prec p\}$ *can be represented in the model MR-Sort if, and only if,* Φ_α^{MIP-D} *is feasible. If the tuple* $\langle \lambda, b, w, c, \mu, \nu, \delta \rangle$ *is a feasible solution of* Φ_α^{MIP-D}*, then the tuple of profiles* b*, the tuple of voting powers* w *and the majority threshold* λ *are suitable parameters of a MR-Sort model that extends the assignment* α*.*

Proof. This theorem results from Theorem 6.1, with only minor changes to the constraints. As noted previously, the normalization constraint (6.1d) has no effect on the feasibility of the problem. Instead, constraints (6.1g) ensure we are looking for voting parameters large enough to have at least a difference of one unit between the votes gathered by any sufficient coalition on the one hand and any insufficient coalition on the other hand. □

6.3 Implementation

In this section, we study the performance of the formulation Φ^{SAT-C} proposed in Section 5.2, both intrinsic and comparative with respect to state-of-the-art techniques. For the implementation, we use a state-of-the-art SAT solver, in order to solve instances of the problem of learning the parameter of a *noncompensatory sorting model with unique set of sufficient coalitions*, as defined in Section 4.1.2 and Section 4.3.2, given the assignment of a set of reference alternatives. We also implement two formulations relying on Mixed Integer Programming, Φ^{MIP-O} and Φ^{MIP-D} presented in Section 6.2, using an adequate solver. We begin by describing our experimental protocol, with some implementation details. Then, we provide the results of the experimental

study concerning the computation time of our program, and particularly the influence the size of the learning set, the number of criteria, and the number of categories, as well as elements of comparison between the three approaches.

6.3.1 Experimental protocol and implementation details

For the experiment we take as an input the assignment of a set of alternatives $\mathbb{X}^\star$, each described by a performance tuple on a set of criteria N, to a set of classes $\{C^1 \prec \cdots \prec C^p\}$.

The performance of the solvers needs to be measured in practice, by solving actual instances of the problem and reporting the computation time required. This experimental study is run on an ordinary laptop with Windows 7 (64 bit) equipped with an Intel Core i7-4600 CPU at 2.1 GHz and 8 GB of RAM.

Dataset generation.

In the scope of this work, we only consider to use a carefully crafted, random dataset as an input. On the one hand, our algorithm is not yet equipped with the capability to deal with noisy inputs, so we do not consider feeding it with actual preference data, such as the one found in preference learning benchmarks [Fürnkranz and Hüllermeier, 2010]. On the other hand, using totally random, unstructured instances makes no sense in the context of algorithmic decision. In order to ensure the preference data we are using makes sense, we use a decision model to generate it, and, in particular, a model compatible with the noncompensatory stance we are postulating. Precisely, we use an MR-Sort model for generating the learning set, a model that particularizes the *noncompensatory sorting model with unique set of sufficient coalitions* by postulating the set of sufficient coalitions possess an additive structure (see Sections 4.3.3 and 6.1.1). This choice ensures the three formulations we are using should succeed in finding the parameters of a model extending the reference assignment.

When generating an instance, we consider the number of criteria $|N|$, the number of ordered categories p, and the number of reference alternatives $|\mathbb{X}^\star|$ as parameters. In order to generate the complete preorders $\langle \succsim_i \rangle_{i \in N}$, we adopt a multiple criteria decision approach, and generate a numeric performance table according to each point of view. We consider all criteria take continuous values in the interval $[0, 1]$, which is computationally more demanding for our algorithm than the case where one criterion has a finite set of values. We generate a set of ascending profiles $\langle b \rangle$ by uniformly sampling $p - 1$ numbers in the interval $[0, 1]$ and sorting them in ascending order, for all criteria. We generate voting weights $\langle w \rangle$ by sampling $|N| - 1$ numbers in

the interval $[0,1]$, sorting them, and using them as the cumulative sum of weights. λ is then randomly chosen with uniform probability in the interval $]0.5,1[$. Finally, we sample uniformly $|\mathbb{X}^\star|$ tuples in $[0,1]^N$, defining the performance table of the reference alternatives, and assign them to categories in $\{C^1 \prec \cdots \prec C^p\}$ according to the model $\mathcal{M}^0$:=MR-Sort$_\omega$, with the parameter $\omega := (\langle b_i^k \rangle_{i \in N, k \in [2..p]}, \langle w_i \rangle_{i \in N}, \lambda)$ grouping the generated profiles, voting weights, and qualified majority threshold.

Solving the SAT problem.

We then proceed by translating the assignment into a binary satisfaction problem, described by sets of variables and clauses, as described by Definition 5.4. This binary satisfaction problem is written in a file, in DIMACS format[c], and passed to a command line SAT solver - CryptoMiniSat 5.0.1 [Soos, 2016], winner of the incremental track at SAT Competition 2016 [d], released under the MIT license. If the solver finds a solution, then it is converted into parameters $(\langle \mathcal{A}^{\text{SAT-C}} \rangle, \mathcal{T}^{\text{SAT-C}})$ for an NCS model with unique set of sufficient coalitions, as described by Theorem 5.3. The model $\mathcal{M}^{\text{SAT-C}}$:=NCS$_{\langle \mathcal{A}^{\text{SAT-C}} \rangle, \mathcal{T}^{\text{SAT-C}}}$ yielded by the program is then validated against the input. As the ground truth $\mathcal{M}^0$ used to seed the assignment is, by construction, an MR-Sort model and therefore a NCS model with unique set of sufficient coalitions, Theorem 5.1 applies and we expect the solver to always find a solution. Moreover, as Theorem 5.3 applies to the solution yielded, we expect the U-NCS model returned by the program should always succeed at extending the assignment provided.

Solving the MIP problems.

We transcribe the problem consisting of finding an MR-Sort model extending the assignment with parameters providing a good contrast into a mixed integer linear optimization problem described extensively in Section 6.2 that we refer to as Φ^{MIP-O}, where O stands for *optimization*. In order to bridge the gap between this optimization stance and the boolean satisfiability approach that is only preoccupied with returning any model that extends the given assignment, we also transcribe the problem consisting of finding *some* MR-Sort model extending the assignment into a MIP feasibility problem (optimizing the null function over an adequate set of constraints), also described in Section 6.2 that we refer to as Φ^{MIP-D}, where D stands for *decision*. These MIP problems are solved with Gurobi 7.02, with factory parameters except for the cap placed

[c] `http://www.satcompetition.org/2009/format-benchmarks2009.html`
[d] `http://baldur.iti.kit.edu/sat-competition-2016/`

on the number of CPU cores devoted to the computation (two), in order to match a similar limitation with the chosen version of the SAT solver. When the solver succeeds in finding a solution before the time limit – set to one hour – the sorting models returned are called $\mathcal{M}^{\text{MIP-O}}$ and $\mathcal{M}^{\text{MIP-D}}$, respectively.

Evaluating the ability of the inferred models to restore the original one.

In order to appreciate how "close" a computed model $\mathcal{M}^c \in \{\mathcal{M}^{\text{SAT-C}}, \mathcal{M}^{\text{MIP-D}}, \mathcal{M}^{\text{MIP-O}}\}$ is to the ground truth $\mathcal{M}^0$ from which the assignment examples were generated, we proceed as follows: we sample a large set of n performance profiles in $\mathbb{X} = [0,1]^N$ and compute the assignment of the corresponding alternatives according to the original and computed MR-Sort models ($\mathcal{M}^0$ and $\mathcal{M}^c$). On this basis, we compute $err - rate$ the proportion of "errors", *i.e.* alternatives which are not assigned to the same category by both models.

6.3.2 Intrinsic performance of the SAT-C formulation

We run the experimental protocol described above by varying the various values of the parameters governing the input. In order to assess the intrinsic performance of our algorithm we consider all the combinations where

- the number of points of view $|N|$ is chosen among $\{5, 7, 9, 11\}$;
- the number of reference alternatives $|\mathbb{X}^\star|$ is chosen among {25, 50, 100, 200, 400};
- the number of categories p is chosen among $\{2, 3\}$.

For each value of the triplet of parameters, we sample 100 MR-Sort models $\mathcal{M}^0$, and record the computation time (t) needed to provide a model $\mathcal{M}^{\text{SAT-C}}$

Figure 6.1 displays the time needed to compute $\mathcal{M}^{SAT-C}$, versus the number of reference alternatives $|\mathbb{X}^\star|$, both represented in logarithmic scale, in various configurations of the number of criteria. The fact that each configuration is seemingly represented by a straight line hints at a linear dependency between $\log t^{SAT-C}$ and $\log |\mathbb{X}^\star|$. The fact that the various straight lines, corresponding to various number of criteria, seem parallel, with a slope close to 1, is compatible with a law where t_{SAT-C} is proportional to $|X^\star|$. The same observations in the plane (number of criteria $\times$ computation time) (not represented) leads to infer a law

$$t^{SAT-C} \propto |\mathbb{X}^\star| \times 2^{|N|},$$

where the computing time is proportional to the number of reference alternatives and to the number of coalitions (corresponding to the number of $|N|$-ary clauses

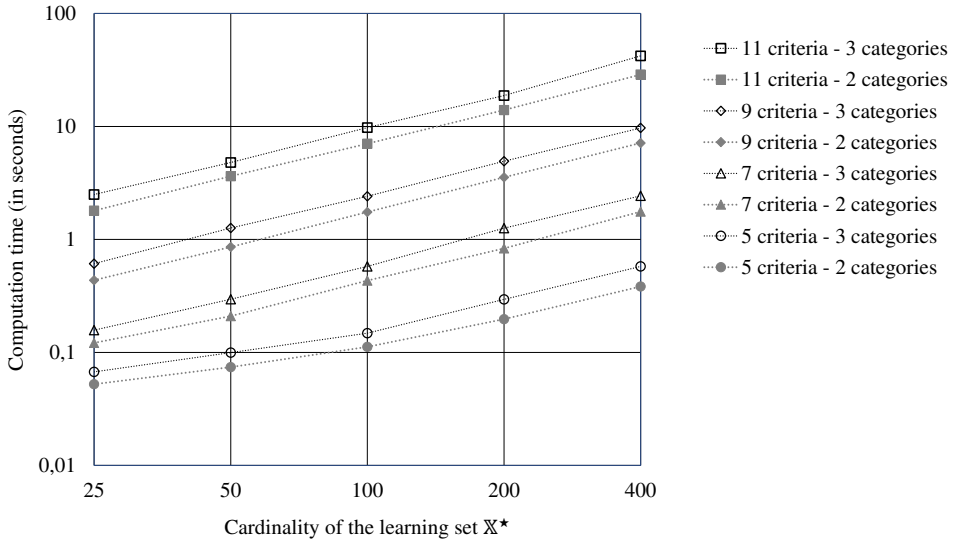

Figure 6.1: Computation time by size of the learning set

of the SAT formulation). Finally, as a rule of thumb: *the average computation time is about 10 s for 11 criteria, 3 categories and 100 reference alternatives; it doubles for each additional criterion, or when the number of reference alternatives doubles.*

6.3.3 Comparison between the formulations

In order to compare between models, we focus on a situation with 3 categories, 9 points of view, and 64 reference alternatives, serving as a baseline. We then consider situations deviating from the baseline on a single parameter – either the number of categories p, from 2 to 5, or the number of points of view, among $\{5, 7, 9, 11, 13\}$, or the number of reference alternatives among $\{16, 32, 64, 128, 256\}$. For each considered value of the triple of parameters, we sample 50 MR-Sort models representing the ground truth $\mathcal{M}^0$, and we record the computation time t needed to provide each of the three models $\mathcal{M}^{\text{SAT-C}}$, $\mathcal{M}^{\text{MIP-D}}$ and $\mathcal{M}^{\text{MIP-O}}$, as well as the generalization indexes for the three models. The MIP are solved with a timeout of one hour.

Results on the computation time.

For the three formulations under scrutiny and the set of considered parameters governing the input, the computation time ranges from below the tenth of a second to an hour (when the timeout is reached), thus covering about five

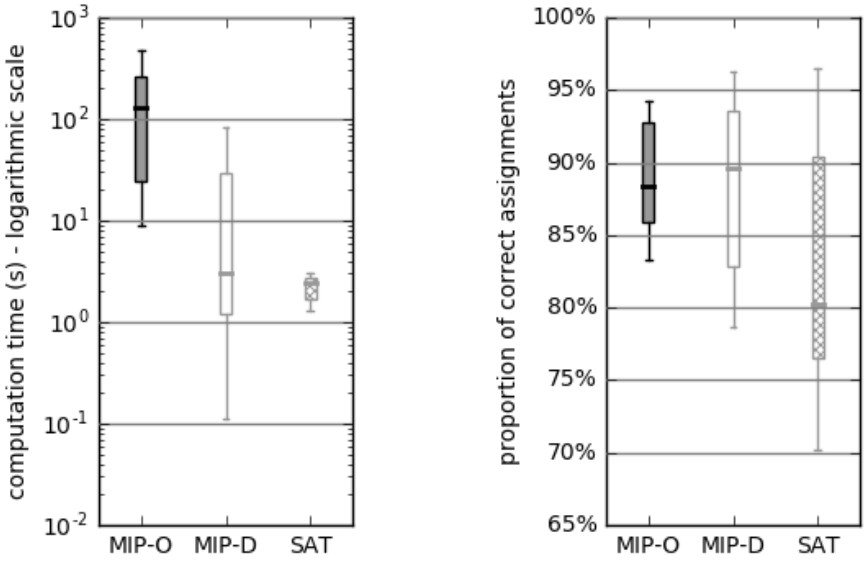

Figure 6.2: Distribution of the computation time and the proportion of assignment similar to the ground truth for the three models in the baseline configuration: 9 points of view, 3 categories, 64 reference alternatives. Represented: median; box: $25-75\%$; whiskers: $10-90\%$.

orders of magnitude. The left side of Figure 6.2 depicts the distribution of the computation time for the baseline situation (9 points of view, 3 categories, 64 reference assignments). While the computing time for the SAT-C and the MIP-D formulations seem to be centered around similar values (with $Med(t^{SAT-C}) \approx 2.4s$ and $Med(t^{MIP-D}) \approx 3.1s$ for the baseline), the distribution of the computing time for the SAT-C algorithm around this center is very tight, while the spread of this distribution for the MIP-D formulation is comparatively huge: the slowest tenth of instances run about a thousand time slower than the quickest tenth. The computation time of the MIP-O formulation appears about 50 times slower than the SAT-C, with a central value of $Med(t^{MIP-O}) \approx 130s$, and covers about two orders of magnitude.

In order to assess the influence of the parameters governing the size and complexity of the input, we explore situations differing from the baseline on a single parameter.

- *The number of reference assignments* $|\mathbb{X}^\star|$. Figure 6.3 indicates that the distribution of the computing time for SAT-based algorithm remains

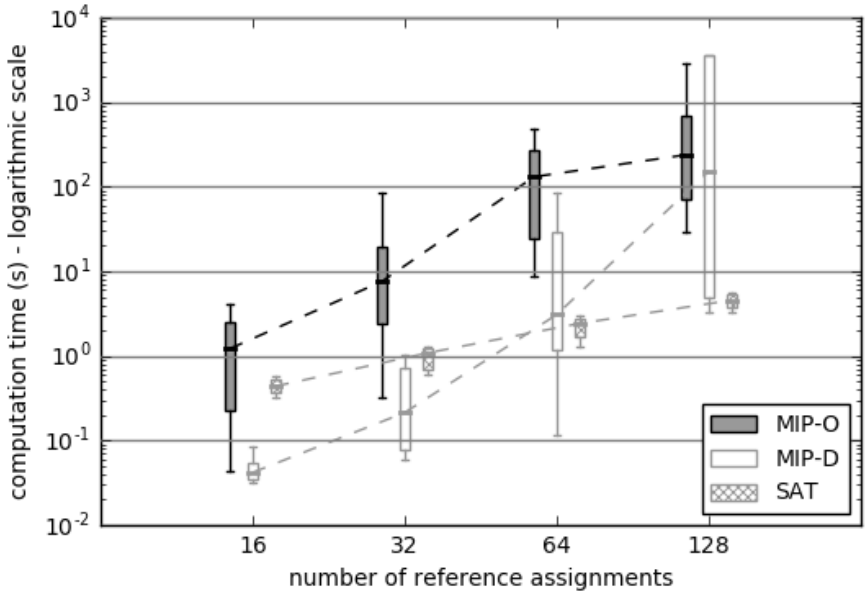

Figure 6.3: Distribution of the computation time for the three models by number of reference assignments, with three classes and nine points of view

tightly grouped around its central value, and that this value steadily increases with the number of reference assignments. Meanwhile, the two MIP formulations display a similar behavior, with an increase of the central tendency steeper than the one displayed by the SAT, and a spread that widens when taking into account additional reference assignments.

- *the number of criteria* $|N|$. Figure 6.4 indicates that the distribution of the computing time for SAT-based algorithm remains tightly grouped around its central value, and that this value steadily increases with the number of criteria. This increase is steeper in the case of the SAT-C and MIP-O formulations than for the MIP-D formulation.

- *the number of categories* p. Figure 6.5 displays an interesting phenomenon. The number of categories seems to have a mild influence on the computation time, without any restriction for the SAT-based algorithm, and as soon as there are three categories or more for the MIP-based algorithm, with a clear exception in the case of two categories, which yields instances of the problem solved ten times faster than with three or more categories.

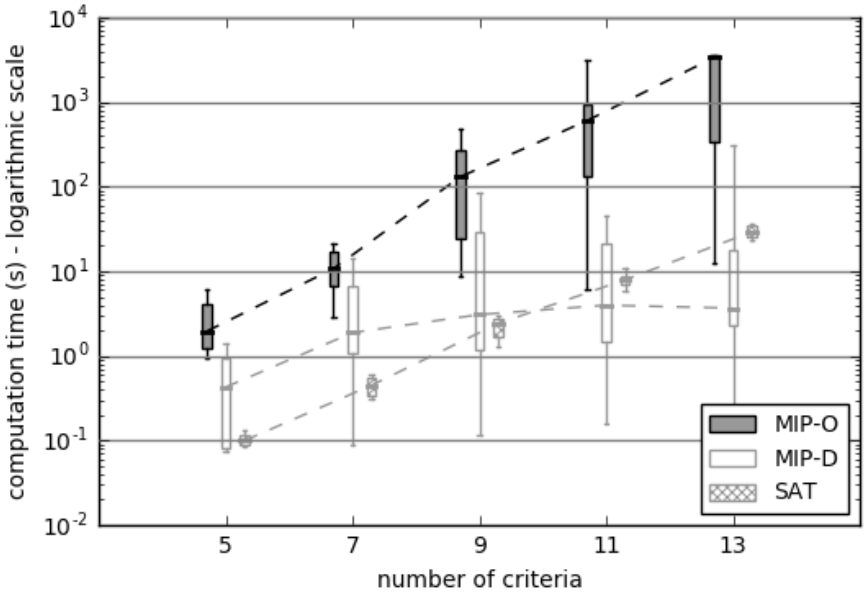

Figure 6.4: Distribution of the computation time for the three models by number of points of view, with three classes and 64 learning examples

Results on the ability of the inferred model to restore the original one.

The right-hand-side of Figure 6.2 depicts the distribution of the proportion of correct assignments (as compared to the ground truth) for the baseline situation (9 points of view, 3 categories, 64 reference assignments). The situation depicted is conveniently described by using the distribution of outcomes yielded by the MIP-D formulation as a pivotal point to which we compare those yielded by the SAT-C and MIP-O formulations: the central 80% of the distribution (between the whiskers) of outcomes for the MIP-O corresponds to the central half (the box) for the MIP-D, while the best half of the distribution of outcomes for the SAT corresponds to the central 80% for the MIP-D. In other terms, compared to the MIP-D, the MIP-O offers consistently good results, while the SAT-C has a 50% chance to yield a model that does not align very well with the ground truth.

Figures 6.6, 6.7 and 6.8 depict the variations of the alignment of the models yielded by the three algorithms with the ground truth with respect to the number of reference assignments, of points of view, or of categories, respectively. The experimental results display a tendency towards a degradation of this

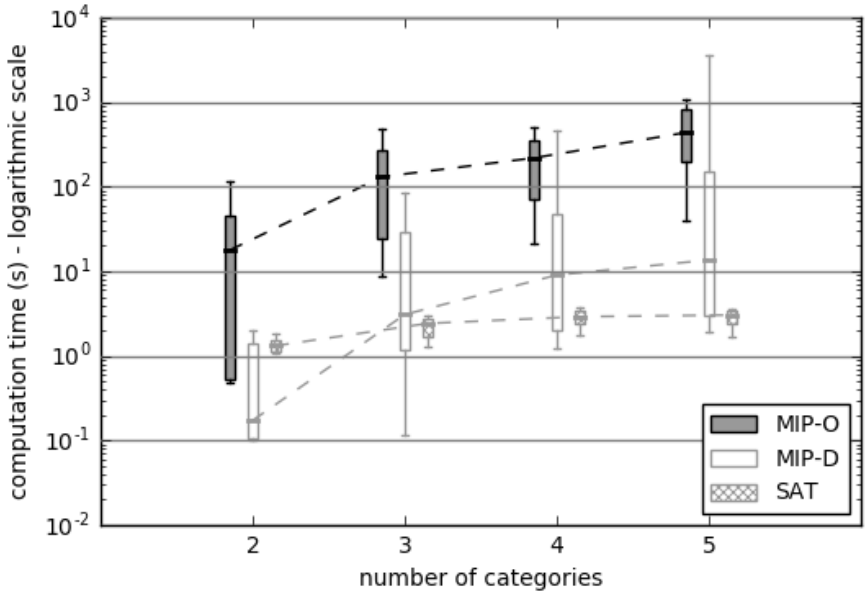

Figure 6.5: Distribution of the computation time for the three models by number of categories, with nine points of view and 64 learning examples

alignment as the number of points of view or the number of categories increase. Conversely, as expected, increasing the number of reference assignments noticeably enhances the restoration rate. The three algorithms seem to behave in a similar manner with respect to the modification of these parameters.

Reliability.

The three formulations expressing the problem we solve—finding a noncompensatory sorting model extending a given assignment of reference alternatives—into technical terms are theoretically faithful. Moreover, as we generate the input assignment with a hidden *ground truth* which itself obeys a noncompensatory sorting model, the search we set out to perform should provably succeed. Unfortunately, a computer program is but a pale reflection of an algorithm, as it is restricted in using finite resources. While we take great care in designing the experimental protocol in order to avoid memory problems, we have purposefully used off-the-shelf software with default settings to solve the formulations. While this attitude has given excellent result for the implementation of the SAT-based algorithm, which has never failed to retrieve a model that succeeds in extending the given assignment, the two MIP-based

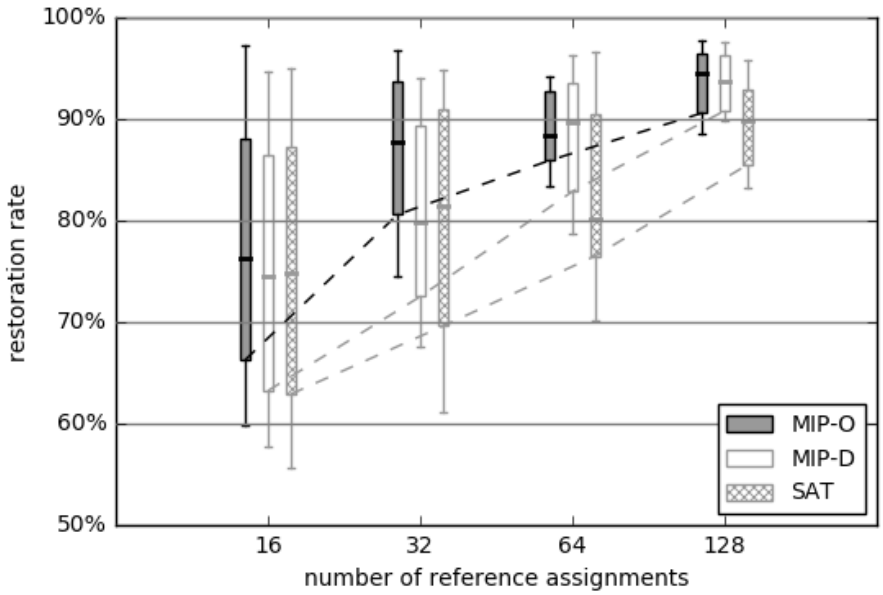

Figure 6.6: Distribution of the generalization index for the three models by size of the learning set, with three classes and nine points of view

implementations have suffered from a variety of failures, either not terminating before the timeout set at one hour or wrongly concluding on the infeasibility of the MIP. We report these abnormal behaviors in Table 6.1.

6.4 Discussion and perspectives

In this section, we strive at interpreting the results presented in Section 6.3. We address the influence of the parameters governing the size and structure of the input - the reference assignment we set out to extend with a noncompensatory sorting model - on the computing time of the programs implementing the three formulations modeling the problem.

6.4.1 Influence of the parameters

The influence of the various parameters ($|X^\star|$, the number of reference assignments; $|N|$, the number of points of view; p, the number of categories) governing the input on the ability of the output model to predict the ground truth seeding the input is best understood from a machine learning perspective. The input assignments form the learning set of the algorithm, while the num-

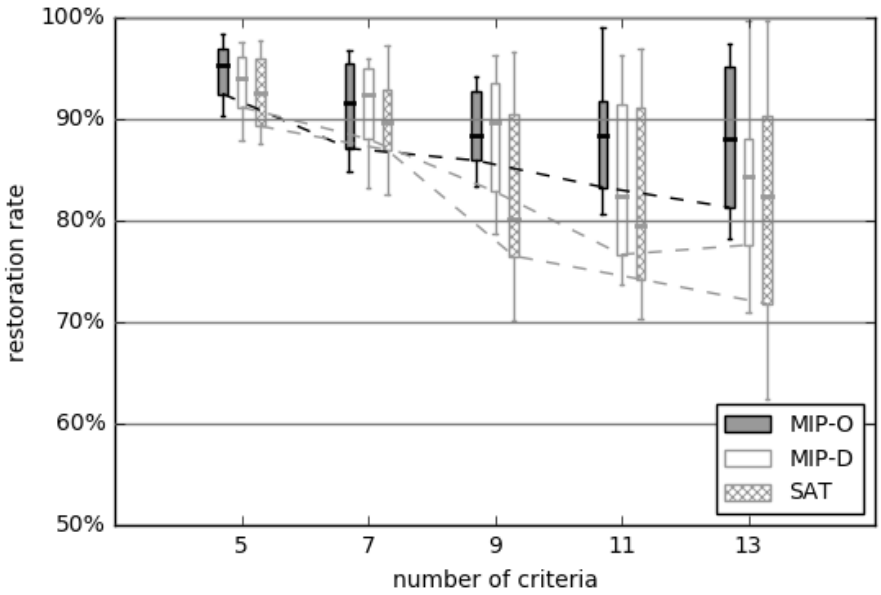

Figure 6.7: Distribution of the generalization index for the three models by number of points of view, with three classes and 64 learning examples

ber of criteria and the number of categories govern the number of parameters describing the noncompensatory sorting model. Hence, an increase in $|\mathbb{X}^\star|$ adds constraints upon the system, while increases in $|N|$ or p relieve some constraints, but demand more resources for their management.

- The comparison between Φ^{MIP-O} and Φ^{MIP-D} informs the influence of the loss function. This influence is threefold: i) optimizing this function demands a lot more time than simply returning the first admissible solution found; ii) formalizing the problem of extending the input assignment with a model as an optimization problem incorporates a kind of robustness into the algorithm, which translates to a decrease in the number of failures; and iii) paradoxically, the strategy consisting in finding the most representative model (in the sense of the chosen loss function) does not yield models with a better alignment to the ground truth than the one consisting to return a random suitable model.
- The MIP-D and SAT-C formulations implement the same binary attitude concerning the suitability of a noncompensatory model to extend a given assignment, and both arbitrarily yield the first-encountered suitable model. Nevertheless, algorithms based on these formulations display

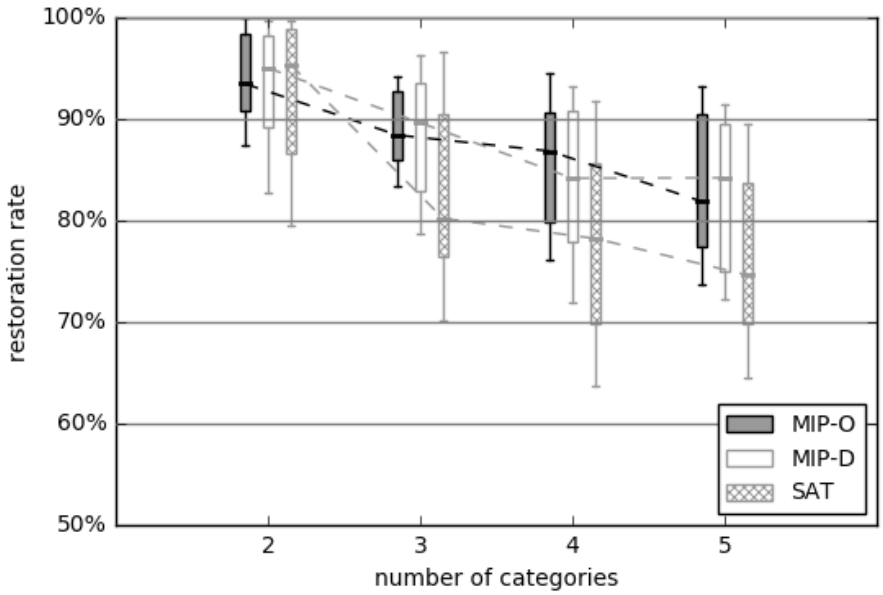

Figure 6.8: Distribution of the generalization index for the three models by number of categories, with nine points of view and 64 learning examples

marked differences in behavior: while the running time of the SAT-based algorithm is very homogeneous between instances and follows very regular patterns when the input parameters change, the MIP-D algorithm behaves a lot more erratically, with some failures (displayed in table 6.1) and a tremendous spread. We credit this difference in behavior to a difference of approach to knowledge representation. Also, with the same input parameters, the model returned by the MIP-D algorithm seems on average to be more faithful to the ground truth than the model returned by the SAT-C algorithm. As both models return random suitable models in different categories (MR-Sort for Φ^{MIP-D}, and the superset NCS with a unique set of sufficient coalitions for Φ^{SAT-C}, while the ground truth is chosen in the MR-Sort category), we interpret the difference in the proportion of correct assignment to the respective volumes of the two categories of model.

- Reference assignments are a necessary evil. On the one hand, they provide the information needed to entrench the model, and refine the precision up to which its parameters can be known. On the other hand, they erect a computational barrier which adds up more quickly for the

$\lvert N\rvert$	5	7	**9**	11	13	9	9	9
p	3	3	**3**	3	3	3	5	7
$\lvert\mathbb{X}^\star\rvert$	64	64	**64**	64	64	128	64	64
MIP-D	4%†	8%†	**4%**	0	0	42%	10%	12%
MIP-O	0	0	**0**	10%	48%	4%	0	0
SAT	0	0	**0**	0	0	0	0	0

Table 6.1: Proportion of instances failing to retrieve a model by size of the input ($|N|$ is the number of points of view, p is the number of categories, $|\mathbb{X}^\star|$ is the number of reference assignment) and formulation. For ease of comparison, the baseline is boldfaced. Failures are usually due to reaching the time limit of one hour, except for configurations marked with a dagger where the failure is due to an alleged infeasibility of the formulation.

MIP formulations we are considering than for SAT-C, as shown in Figure 6.3. Overcoming this barrier demands time and threatens the integrity of the somewhat brittle numerical representation underlying Φ^{MIP-D}.

- From the perspective of the model-fitting algorithm, the number of points of view and the number of categories are usually exogenous parameters, fixed according to the needs of the decision situation. The specific numbers of points of view we considered during the experiment, from 5 to 13, cover most of the typical decision situations considered in Multiple Criteria Decision Aiding. Introducing more points of view demands to assess more parameters, which has a compound effect on complexity, as it requires at the same time to build a higher-dimension representation of the models, and to provide more reference examples in order to be determined with a precision suitable to decision making. Apart from a noticeable exception (see below), the number of categories does not seem to have much influence (as shown on figures 6.5 and 6.8).
- Underconstrained models are not very good at providing recommendations. When fed with scarce information, the task of finding a suitable extension is easy, but there are very little guarantees this extension matches the unexpressed knowledge and preferences of the decision maker concerning alternatives outside the learning set. We interpret the decrease in the ability to align with the ground truth as the number of criteria increases displayed on Figure 6.7 as an expression of an *overfitting* phenomenon, where too many parameters are chosen to faithfully represent a too little slice of the set of alternatives, but poorly represent cases never seen before.

- Mixed integer programs can represent decision problems, in theory. Practically though, some complex inputs have proven overwhelming for the MIP-D formulation, whereas the MIP-O has shown more robustness, as evidenced by Table 6.1. It seems fair to assume this lack of stability is related to the absence of a normalization constraint such as (6.1d) in Φ^{MIP-D} formulation . Determining a good lower bound on the difference of normalized voting power between sufficient and insufficient coalitions would therefore likely help alleviating this issue.
- MR-Sort with two categories is structurally different than models with more than two categories. While we have defined it as a procedure where alternatives are compared holistically to a profile, it can also be described as an additive value sorting model with stepwise, non-decreasing, 2-valued marginals. The experimental results, both for the computing time and the alignment with the ground truth (see figures 6.5 and 6.8, where the points corresponding to two categories are outliers with respect to the rest of the series) highlight this peculiarity, and tend to show that the value-based representation of the MR-Sort model with two categories is computationally efficient.

See also

Chapter 5 details two formulations based on Boolean satisfiability in order to solve the inverse noncompensatory sorting problem, but this chapter only investigates the performance of one of them, SAT-C. Are we even sure the second one, SAT-P, is correct? How do the two formulations compare to each other?

[Belahcene et al., 2018b] addresses this issue, and is reproduced in Appendix C.

7

Accountable decisions with Inv-NCS

The previous chapters of Part II were devoted to the definition and the solving of the Inv-NCS problem: given points of view expressing preferences and an assignment of alternatives to ordered categories, is it possible to represent this assignment in the noncompensatory sorting model? if possible, with which parameter? This chapter addresses the question of the use of this new tool in the context of decision aiding situation. In Section 7.1, we consider a *robust elicitation* process, and question the contribution of Inv-NCS to this process. In Section 7.2, we address a fictitious decision situation, where the process is subject to seemingly contradictory accountability demands of transparency and secrecy.

7.1 Robust elicitation of a noncompensatory sorting model

7.1.1 The situation

As detailed in Section 1.1, we consider a *decision aiding process*, as described in [Bouyssou et al., 2006, Tsoukiàs, 2008], occurring between somebody looking for decision support, and an analyst providing such a support.

Additionally , we assume that:

i) alternatives need to be assigned to categories, where the set of possible categories is known, finite and totally ordered;

ii) for the aggregation step, the decision maker and the analyst have agreed upon the relevance of using a noncompensatory procedure, maybe for one or more reason among those listed in the introduction of Part II.

Examples of real-world applications of decision aiding processes following this assumption can be found in [Figueira et al., 2005], modeled after an Electre Tri procedure [Mousseau et al., 2000], and in [Bisdorff et al., 2015], modeled after an MR-Sort procedure [Leroy et al., 2011]. For our concern, we assume the aggregation is following the noncompensatory sorting model [Bouyssou and Marchant, 2007a, Bouyssou and Marchant, 2007b], as detailed in Chapter 4, in its 'vanilla' version, i.e. without considering any of the variants listed in Section 4.3.

Therefore, according to [Bouyssou et al., 2006], what remains to be done consists in aggregating the evaluations corresponding to the different point of view, and deriving a final recommendation. Section 7.2 proposes to jointly consider the aggregation step with the final recommendation, by anticipating external demands of *accountability* of the recommendation. In this section, we mostly focus on the aggregation step, and *accountability* is considered inside the scope of the decision aiding process—the dialog between the decision maker and the analyst.

7.1.2 State-of-the-art approaches

Opting for the noncompensatory sorting model is a key step for the process of aggregating the array of evaluations made according to the various points of view into an overall sorting of the candidates into the ordered categories. Nevertheless, at this stage, this option only translates into an empty framework. The parameter governing the model—approved sets according to each point of view and at each level of requirement, sufficient coalitions of points of view at each level of requirement—still needs to be *elicited* in order to describe a *sorting procedure* mapping alternatives to categories. How should this parameter be determined? The stake is to reflect, as precisely as possible, the personal judgment of the decision maker, by tuning the preference parameter. Several approaches to this question have been proposed, detailed, and implemented:

Direct elicitation: it can be argued that the preference parameter of the noncompensatory sorting model is readily *interpretable* by the decision maker. They directly reflect, at a given level of requirement, the expectations of the decision maker—as opposed, e.g. of the parameters of the additive model, even the weighted sum, that only reflect acceptable trade-offs through their ratios. It seems reasonable to expect that, in some situations, the preference parameter

would directly be set by the decision maker and the analyst, maybe using some trial and error, through simulation tools, to check the relevance of this setting w.r.t. the output of the model. Nevertheless, when the number of points of view increases, the explicit representation of coalitions of criteria becomes exponential and therefore difficult to handle, visualize, and maybe understand. In order to circumvent this source of complexity, the popular variants Electre Tri and MR-Sort propose to represent the power of coalitions of points of view with an additive model. While easy to handle, it has been argued that this representation might induce confusion and fallacious impressions in the mind of the decision maker[a]. Consider, for instance, the following example.

Example 7.1. As an example, consider the two sets of weights and majority threshold given in Table 7.1, such that a coalition $B \subset \{1, 2, 3\}$ is considered sufficient if, and only if, $\sum_{i \in B} w_i \geq \lambda$.

w_1	w_2	w_3	λ
49	11	40	50
1	1	1	2

Table 7.1: Two sets of parameters—voting powers and qualification threshold—representing the same sufficient coalitions of points of view.

Both sets lead to the same winning coalitions: a combination of at least two of the three criteria is needed in order to be above the majority threshold. The first set of parameters lets the decision maker think that there is an order of importance between the three criteria since $w_1 > w_3 > w_2$. This insight is not true when one looks at the list of winning coalitions: each criterion has the same importance. The second set of parameters does not let appear such a hierarchy between the weights.

Indeed, different sets of weights and majority threshold can represent the same sets of winning coalitions. The value of these weights and majority threshold can be misleading. The claims of *interpretability* of the model should be backed up by a careful examination of the preference parameters under the lens of *conjoint measurement*: maybe a parameter offers a convenient way of dialing the model and making sure it fits the attitude of the decision maker, and maybe the procedures following the model are easy to carry out, but what cannot be measured should be interpreted with the greatest care.

[a] And maybe in the mind of some analysts too!

Full indirect elicitation via standard sequences. The preference parameter describing the personal preferences of the decision maker can be obtained by asking them questions about the category where fictitious, carefully chosen alternatives should be assigned:

- In a first phase, questions aim at pinpointing the set of sufficient coalitions. They involve fictitious alternatives ideally fit according to some points of view, and absolutely unfit according to the rest.
- A second phase is devoted to finding the acceptance threshold according to each criterion. This information can be obtained e.g. by degrading the fitness on a particular criterion $i \in N$, starting from a fictitious alternative that is ideal on every criteria of a minimally sufficient coalition containing i, and absolutely unfit according to the points of view outside this coalition.

This elicitation procedure involves a lot of difficult questions concerning fictitious alternatives that combine extremely bad and extremely good attributes and border the bizarre. It might be feared that, by the simple virtue of being so remote from reality:

- these alternatives are perceived as toys, in which the decision maker does not believe. Therefore, their assessment has little counterfactual value.
- thinking about these alternatives removes the decision maker from the situation at hand in an alienating manner.

Learning from assignment examples. The full indirect elicitation, described above, empowers the analyst with the entire responsibility of the acquisition of preference information, without any restriction w.r.t. the alternatives presented to the decision maker. A more balanced approach with respect to the roles and responsibilities of the decision maker and the analyst during the decision aiding process may lead to consider preference information concerning alternatives that are not crafted by the analyst, then fit the parameters in order to 'best' represent these preferences inside the model. For sorting problems, UTADIS [Devaud et al., 1980] is a pioneering implementation of this idea, assuming an additive value model. This approach sometimes called *disaggregation* [Jacquet-Lagrèze and Siskos, 2001, Zopounidis and Doumpos, 2002, Mousseau, 2003] can be described as a four-step process, detailed in Table 1.1.

The disaggregation framework has proven itself, time and time again, with additive value models (for sorting or comparing): with such a model, the second step yields a single linear comparison representing a statement of the preference information in the parameter space. Consequently, the set

of parameters compatible with the preference information is a polytope, and the third step can be resolved using linear programming, as soon as the loss function is chosen linear. As an added substantial benefit, the restitution of the preference statements can be integrated into the loss function, rather than with the frontiers of the polytope. This Lagrangian technique, often dubbed 'soft constraints', enables this framework to seamlessly deal with inconsistent preference information—cases where the polytope of compatible parameters is empty.

When trying to apply the disaggregation framework to any variant of the noncompensatory sorting model, difficulties arise:

- The transcription of preference information statements into constraints on the preference parameter is cumbersome. Proposed transcriptions rely on nonlinear constraints [Mousseau and Slowinski, 1998] for Electre Tri or mixed integer programming [Leroy et al., 2011, Sobrie et al., 2013] for MR-Sort. In particular, no representation under the form of linear comparison over continuous variables has been proposed.

- As a consequence of the previous point, computation of an optimal, or even suitable, parameter is difficult, and painfully slow. Solving an instance with seven points of view takes minutes [Leroy et al., 2011].

In order to circumvent these obstacles, [Sobrie et al., 2015] proposed to learn the parameter of an MR-Sort model from assignment examples using a heuristic method based on a population of parameter values. This algorithm converges quite rapidly towards practically good solutions, but without offering any guarantee.

7.1.3 Contributions of Inv-NCS

Inside the disaggregation framework. Chapter 5 introduces a novel approach to the second step of the disaggregation framework for sorting described by Table 1.1 for the noncompensatory sorting model. Instead of relying on sophisticated languages such as nonlinear programming or mixed integer programming, constraints are expressed in the elementary language of propositional logic, as a disjunction of atoms. Parameters permitting to restore the preference information correspond to solutions of a Boolean satisfiability problem.

At this stage, a first important contribution is a proof of the NP-hardness of solving the Inv-NCS problem. Unless $P = NP$, it entails the impossibility to express the constraints reflecting the preference information under the form of linear comparisons over continuous variable.

In Chapter 6, either one or the other of the two Boolean satisfiability (SAT) formulations of Inv-NCS are used to power the second and third steps of the disaggregation framework. It is an actual departure of the usual approach though, as the SAT paradigm, in its simplest form[b], is solely concerned about the feasibility of a constrained problem, but is alien to the notion of optimization. Therefore, in the third step, the derived value of the parameter is arbitrary, corresponding to the first solution encountered by an external solver. A thorough experimental validation allows to derive the following findings:

- *Speed*—the proposed SAT-based architecture permits to obtain results about fifty times more rapidly than the state-of-the-art MIP-based one. Hours of computation time become minutes, minutes become seconds, and this massive increase in performance opens new perspectives for the usage of noncompensatory sorting models.
- *Stability*—the proposed SAT-based architecture is predictable in its efficiency, and offers reliable results, that can be trusted. That was not the case with the MIP-based architecture, as the representation of an intrinsically discrete parameter space by means of continuous variables naturally entails stability issues at the frontiers, and delegating these issues to the solvers does not offer solid guarantees.
- *Relevance*—there is two aspects to this notion: i) is it wise to move away from representing sufficient coalitions with the majority rule? and ii) is it wise to do without a loss function? To the first question, the experimental results are not sufficient to assert a positive answer, but they nevertheless establish our option is certainly not foolish. NCS is a broader model than MR-Sort. This additional breadth does not incur an additional computational cost—quite the contrary—but it certainly incurs an additional informational cost—the need to obtain more preference information in order to pinpoint the preference parameter—that cannot yet be precisely estimated. Concerning the loss function, it might be practically difficult to do without—and this aspect requires further investigation—but using one to infer a precise preference parameter poses a problem of *accountability*, specific to the noncompensatory sorting model, that we discuss in the next paragraph.

Robust disaggregation. Finding a suitable parameter for the NCS problem is literally a byproduct of the Boolean satisfiability formulations of Inv-NCS. Their primary focus is on the *feasibility* of this search. This feature can readily

[b] We address the opportunity of relying on more sophisticated languages, such as MaxSAT, in Chapter 8.

be leveraged, via Definition 7.1 into the *robust disaggregation* framework, a variant of the disaggregation framework described in [Salo and Hamalainen, 2001, Greco et al., 2008, Greco et al., 2010b], and summarized in Table 1.2. It relies on the following definition, tying the *necessary and possible* recommendations for the assignment of a candidate an the feasibility of the inverse problem:

Definition 7.1. *Given some preference information $\mathcal{P}$, a model $\mathcal{M}$ consistent with $\mathcal{P}$, and an output π in the codomain of $\mathcal{M}$:*

- *we say π is* impossible *when $\mathcal{P} \cup \{\pi\}$ is inconsistent with $\mathcal{M}$, i.e. $\mathcal{P} \cup \{\pi\}$ is a negative instance of the inverse-$\mathcal{M}$ problem;*
- *we say π is* necessary *when every other output in the codomain of $\mathcal{M}$ is impossible.*

The robustness offered by this framework is epistemic by nature: it protects the recommendation from becoming irrelevant, should additional preference information become available. The price to pay for this insurance against our own ignorance is obviously quite high.

- in terms of *computational burden*—instead of simply applying the model, which is really easy in the case of NCS, it requires to solve the NP-complete Inv-NCS problem for each query;
- in terms of *simplicity*—a parameterized NCS model is arguably easy to interpret for the decision maker, as it relies on a compact narrative about sufficient fitness according to each point of view, and sufficient coalitions of points of view, and the robust version does not offer such convenient metaphors;
- the robust model is irresolute—the *possible* assignments for a given candidate alternative form a subset of the categories, that decreases when the preference information increases, but it sometimes does not boil down to a singleton (a *necessary* assignment).

The issue of *simplicity* is particularly acute in the case of noncompensatory models. Compare the situation to the additive value model, when preference information is a set of pairwise comparative statements, as detailed in Part I. Any pair of alternatives $(x, y) \in \mathbb{X}^2$ partitions the parameter space into two half-spaces separated by a hyperplane, and the compatible preference parameters reside in a polytope. The size of this polytope can be meaningfully measured in terms of maximal pairwise regret [Wang and Boutilier, 2003], and the obtaining of additional preference information should aim at reducing this size. When the polytope of compatible parameters is small enough, it might reasonably well be approximated by a ball. A ball corresponds to another paradigm

permitting to deal with uncertainty, with a central tendency and a scalar measure of residual uncertainty corresponding respectively to the center and the radius of the ball. When preference information is scarce, the (deductive) polytope has few facets—one per pairwise comparative statement— and might not be well approximated by a ball. With abundant preference information, the representation of the polytope with constraints becomes costly, and the polytope shrinks, so the loss due to replacing the polytope with a ball becomes minor. Also, when the polytope becomes small, the specifics of the inference procedure—which consists in selecting a single parameter inside the polytope—becomes irrelevant, as all compatible parameters are close to each other. In the case of noncompensatory sorting models, the situation is quite different. For instance, when the attribute scales are continuous, MR-Sort can be described with continuous parameters—limiting profiles, voting powers and a majority threshold. Inside this continuous space, though, the subset of parameters compatible with a given assignment of reference alternatives is likely *not* a polytope, and it is probably not even connected. This is because of the nonlinear interaction between limiting profiles and voting powers: the same situation is likely to be represented with either narrow accepted sets and a wide set of sufficient coalitions, or the converse. An inference procedure—for instance, one corresponding to a loss function—is asked to select a representative parameter. With a fragmented subset of compatible parameters, this choice is necessarily:

i) arbitrary; and

ii) potentially insincere, because it mimics a procedure which is totally acceptable in the linear case mentioned previously.

Therefore, while robust decision aiding might be considered a nice-to-have, but costly, tool in the case of linear models, its role becomes crucial when considering to provide *accountable* recommendations supported by a nonlinear model.

Interlude

- In the previous section, the version space of the NCS model reflected the incomplete knowledge of the analyst concerning the preferences of the decision maker, and represented the envelope of possibly preferred outcomes of the sorting models. The inverse noncompensatory sorting problem can be used as a tool for guaranteeing the sincerity of the elicitation process, by correctly aligning the views of the decision maker, of the analyst, and the actual product of the process.
- In the following section, the version space of the NCS model is considered as the leeway left to a jury in its appreciation of candidates that ought

to be sorted into ordered categories. This autonomy left to the jury needs to strike a delicate balance: should the model be entirely specified, the jury could be fully automated, and its functioning could be gamed and manipulated by malicious stakeholders; on the opposite end of the spectrum, a very loose specification of the model results in a practical lack of control of the jury by the society, and fails to prevent possible abuses of authority. Inv-NCS is used as a tool to account for the decision of the jury, given some specifications.

7.2 Accountable Approval Sorting

A committee meets to decide upon the sorting of a number of candidates into two categories (e.g. candidates to accept or not, projects to fund or not). The committee applies a decision process which is public, the outcomes are public as well, however the details of the votes are sensitive and should not be made available. To what extent can we make the committee accountable of his decisions?

Most of the material presented in this section was originally published in [Belahcene et al., 2018a], together with the pairwise characterization of positive instances Inv-NCS exposed in Section 5.3.

7.2.1 The context: selection by a jury

We shall primarily be concerned with a general sorting model where candidates are sorted by a jury N. Each juror $\text{🧍}_x \in N$ expresses binary judgments [Laslier and Sanver, 2010], and candidates are sorted as either *good* or *bad* depending on the fact that the coalition of jurors supporting this sorting is strong enough, or not, to win the decision of the jury.

Example 7.2. 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:

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

The notions involved in this real-life situations map straightforwardly to the primitives of the noncompensatory sorting model: candidates are *alternatives*, jurors are *points of view*, and we are considering two *categories* {Bad ≺ Good}. For the noncompensatory sorting 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 any dynamics in the way each juror appreciates the candidates. This preference may stem from numeric or symbolic performance, as it is often the case in multiple criteria decision aiding, or be intrinsically ordinal, as it is often assumed in social choice contexts.
- *Indivividual consistency between preferences and vote.* Each juror $\text{⚲}_{\text{✕}} \in N$ is allowed to express only a binary judgment on each candidate $x \in \mathbb{X}$, which is either 'approved according to $\text{⚲}_{\text{✕}}$' or not. The approved subset of candidates $\mathcal{A}_{\text{⚲}_{\text{✕}}} \subseteq \mathbb{X}$ should be an upset for the preference relation $\succsim_{\text{⚲}_{\text{✕}}}$. Hence, there is no pair of candidates $x, x' \in \mathbb{X}$ where x is preferred to x' w.r.t. $\succsim_{\text{⚲}_{\text{✕}}}$, x' is approved by $\text{⚲}_{\text{✕}}$ 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—it can be described, in terms of cooperative game theory, as a *monotonic simple game* [Peleg, 2002]. In particular, this precludes the additive structure assumed in *weighted majority games* [Peleg, 2002][c] or *approval balloting* [Laslier and Sanver, 2010]. The jury has to find a 'consensual agreement', but what defines an acceptable consensus remains unclear—from the public, if not for the jurors themselves. Indeed, the notion of 'sufficient coalitions' may appear unwieldy to use explicitly, as it references the power set of the points of view.

Example 7.3. (Example 7.2 continued) Suppose the approved sets are as follows: $\mathcal{A}_{\text{⚲}_{\text{✕}}^1} := \{a, b, f\}, \mathcal{A}_{\text{⚲}_{\text{✕}}^2} := \{e, b, c\}, \mathcal{A}_{\text{⚲}_{\text{✕}}^3} := \{f, a, b\}, \mathcal{A}_{\text{⚲}_{\text{✕}}^4} := \{d, a, c\},$

[c] We note that there are simple voting structures, such as e.g. *bicameralism*, that cannot be represented as weighted majority games.

$\mathcal{A}_{\text{♀}^5} := \{c, e, b\}$, 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 noncompensatory model assigns a, b, c to the Good category, and d, e, f to the Bad one. Hence, $\alpha := \{(a, \text{Good}), (b, \text{Good}), (c, \text{Good}), (d, \text{Bad}), (e, \text{Bad}), (f, \text{Bad})\}$. We note the same assignment α can be obtained with different sorting parameters, e.g. approved sets $\mathcal{A}'_{\text{♀}^1} := \{a, b, f\}, \mathcal{A}'_{\text{♀}^2} := \{e, b, c, d, a\}, \mathcal{A}'_{\text{♀}^3} := \{\}, \mathcal{A}'_{\text{♀}^4} := \{d, a, c\}, \mathcal{A}'_{\text{♀}^5} := \{c\}$ and sufficient coalitions $\mathcal{S}'$ containing the coalitions $\{1, 2\}, \{5\}$ and their supersets.

7.2.2 Accountability requirements

While the jury as a whole has the power of taking 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[d]—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, Pettit, 2000] is contestability: a democratic institutional arrangement should be such that citizens could effectively challenge public decisions. The control of the governed on the government is in general 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 of the process regarding his own classification, could challenge the committee and asks for a justification.

This 'right for an explanation', put forward with the GDPR at the level of the European Union, is actually entrenched in the French legal tradition under the form of an obligation, for any administration, to motivate its decisions and, more generally, from the principle of contradictory debate. We believe, as [Kroll et al., 2017] or [Cozic and Valarcher, 2017] that institutions should be organized so as to make contestation possible and that this requirement should be taken into account from the inception of the procedure (and carried forward throughout its life cycle).

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

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*, an ill-defined notion in general, but that we use here as meaning 'comprehensible by the persons that need to'—here, the contestant. In the context of jury decisions, transparency is out of 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 authority of the jury. In France, for instance, it is customary, and even mandatory for justice decisions, that the deliberation of a jury remains secret, and that the final decision is attributed to the jury—or the Court—as a whole. The consensus between jurors—or judges—has to be found beforehand, in the privacy of the deliberation, and there is no place left for dissenting opinions when the decision is made public. This process contrasts with the anglo-saxon tradition of publishing the decisions of each judge.
- Revealing the decision rule, or publishing a lot of information about it, would create a feedback effect with some candidates adopting a strategic behavior in order to game the output. This behavior might be itself detrimental to the goals of the jury, and might also be detrimental to other candidates less knowledgeable about the system, creating a breach in equity.

Complexity Leaving the burden of proof on the shoulders of the audit agency, or worse, of a lone plaintiff, may be too demanding. It requires, at

the same time, to give access to a lot of information—possibly the preferences and the assignment of the whole set of candidates—and to solve difficult combinatorial problems—we have shown that the Inv-NCS problem is NP-hard (Theorem 5.11)—that scales badly with the number of candidates.

Sections 7.2.3 and 7.2.4 respectively focus on answering the procedural regularity and the contestability requirements, while paying attention to disclose as few information as necessary, and providing explanations that are comprehensible by their recipient.

7.2.3 Addressing overall *Procedural regularity* with Inv-NCS

In this section, we consider how participants, decision subjects, and observers can be assured that each individual sorting decision was made according to the same procedure——for example, how observers can be assured that the jury is not choosing outcomes on a whim while merely claiming to follow an announced rule. In particular, we assume this task is delegated to an independent audit agency, which is in charge of overseeing the whole process on behalf of the stakeholders of the decision.

In view of this demand, what needs to be proven is that α is a positive instance for the Inv-NCS problem, i.e. the assignment α is a *possible* outcome for NCS, given the preferences of the jurors over the candidates.

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

i) full disclosure of the preference profile $\langle(\mathbb{X}, \succsim_i)\rangle_{i\in N}$, and the auditor solving the NP-hard Inv-NCS problem, e.g. using a SAT solver and either of the formulations Φ_α^{SAT-C} or Φ_α^{SAT-P} detailed in Chapter 5; 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 by Lemma 5.5.

Note that the full disclosure of the decision rule is not even on the table. It would require to reveal 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 a 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 5.6 to compute and

provide a certificate of feasibility for Inv-NCS(α) that involves the disclosure of less information, as illustrated below:

Example 7.4. (Example 7.3 continued) 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 in Section 7.2.1and 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 can be formalized into an *argument scheme*, an operator tying a tuple of premises – pieces of information satisfying some conditions – to a conclusion [Walton, 1996].

Definition 7.2 (Argument Scheme (AS1))**.** *We say a tuple* $\langle (i_1, g_1, G_1, b_1, B_1), \ldots, (i_n, g_n, G_n, b_n, B_n)\rangle$ *instantiates the argument scheme AS1 supporting the assignment* α *if: i) for all* $k \in \{1 \ldots n\}$, $i_k \in N$, $g_k \in G_k$, $\alpha(G_k) = \{\textsc{Good}\}$, $\forall g \in G_k, g \succsim_{i_k} g_k$, $b_k \in B_k$, $\alpha(B_k) = \{\textsc{Bad}\}$, $\forall b \in B_k, b_k \succsim_{i_k} b$ *and* $g_k \succ_{i_k} b_k$*; and ii)* $\bigcup_{k\in\{1\ldots n\}} G_k \times B_k = \alpha^{-1}(\textsc{Good}) \times \alpha^{-1}(\textsc{Bad})$

Hence, according to the point of view i_k, g_k is the least preferred alternative in the subset of Good alternatives G_k and it is preferred to b_k, the most preferred alternative in the subset of Bad alternatives B_k. This scheme is somewhat frugal in the number of pairs of the profile $\langle(\mathbb{X}, \succsim_i)\rangle_{i\in N}$ revealed to the auditor, as the comparisons inside $G_k \times G_k$ or $B_k \times B_k$ are not disclosed. Theorem 5.6 can be reworded as follows:

Corollary 7.1. *An assignment* α *is a positive instance of Inv-NCS if, and only if, there is an instance of AS1 supporting it.*

Example 7.5. (Example 7.4 continued) The explanations given in Example 7.4 instantiate AS1 as follows: $\big\langle(1, b, \{a, b\}, e, \{d, e\}), (2, c, \{b, c\}, d, \{d, f\}), (4, c, \{a, c\}, e, \{e, f\})\big\rangle$

The shift in the burden of proof allows the jury to support its claim (here, the result of the sorting procedure) with arguments of its own choosing. The length n of an explanation instantiating the argument scheme AS1 offers an indication regarding its cognitive complexity as well as the amount of information disclosed to the auditor. Therefore, we would rather provide the shortest possible explanations, and strive to mention as few points of view as possible. Obviously, an explanation needs to reference a specific point of view at most once, so $n \leq |N|$. Unfortunately, the following result shows that one might require all points of view in a complete explanation, even in situations with relatively few alternatives.

Proposition 7.2. *For every jury* N*, there exists a set of* $|N|+1$ *alternatives* $\mathbb{X}$ *and an assignment* $\alpha : \mathbb{X} \to \{\textsc{Bad} \prec \textsc{Good}\}$ *for which any tuple instantiating the argument scheme AS1 and supporting* α *has length* $|N|$*.*

Proof. The result is shown by induction on the size of the jury n, with n candidates assigned to Good and one assigned to Bad.

- *Base case:* for $n = 1$, i.e. $N = \{\mathbf{\Omega}^1\}$, we consider the candidates $\mathbb{X}_2 = \{b, g\}$, the preference $g \succ_{\mathbf{\Omega}^1} b$ and the assignment $\alpha_1 := \{(g, \textsc{Good}), (b, \textsc{Bad})\}$.
- *Inductive step:* Consider an assignment α_n on $n+1$ candidates $\mathbb{X}_{n+1}$, with $(b, \textsc{Bad}) \in \alpha$ assessed by the jury of size n: $N = \{\mathbf{\Omega}^1, \ldots, \mathbf{\Omega}^n\}$, but no strictly smaller jury. We introduce a new candidate z, and a new juror $\mathbf{\Omega}^{n+1}$. Then: the jury $N \cup \{\mathbf{\Omega}^{n+1}\}$, with preferences such that:
  - $\mathbf{\Omega}^{n+1}$ is indifferent w.r.t. all candidates in $\mathbb{X}_{n+1}$, and prefers z to all of them;
  - all the jurors in N prefer z to some candidate $b \in \alpha_n^{-1}(\textsc{Bad})$.

  the alternatives $\mathbb{X}_{n+2} := \mathbb{X}_{n+1} \cup \{z\}$ and the assignment $\alpha_{n+1} := \alpha_n \cup \{(z, \textsc{Good})\}$ satisfy the property at the step $n+1$: any tuple instantiating the argument scheme AS1 and supporting α_{n+1} has length $n+1$. Indeed, it has to take into account the juror $\mathbf{\Omega}^{n+1}$ to positively discriminate z from b; but this juror does not help to discriminate any pair of candidates in α_n.

□

7.2.4 Contestability of individual decisions

In this section, we focus on the situation where a candidate, supposedly unsatisfied with the outcome of the decision process regarding its own classification, challenges the committee and ask for a justification.

Explaining the outcome of a classifier. This question can be considered as falling under the umbrella of 'explaining the outcome of a classifier', that has fostered a lot of interest for the last thirty years. The problem of devising relevant explanations accompanying the result of a classification algorithm has been addressed from many angles, e.g.:

a. *Interpretation:* going step by step through the classifier has reached the particular conclusion concerning this candidate. This approach has risen with the advent of rule-based system, a.k.a. *expert systems* [Buchanan and Shortliffe, 1984, Waterman, 1986], where the trace of the algorithm corresponds to the triggering of a number of business rules that were thought to make for an explanation. Limits of this approach have been discussed, e.g. [Alvarez, 2004], pointing at the fact that a particular trace of the execution is often arbitrary and might be misleading. In our case, the noncompensatory sorting model has the advantage of being easy to interpret: a candidate is deemed Bad if, and only if, it has been disapproved by the complementary of a sufficient coalition of jurors.

b. *Surrogation:* For complex classifiers, finding a surrogate model that can be interpreted: in order to explain a recommendation made about a particular candidate by the model, a so-called *interpretable* model is learned in the neighborhood of the candidate. This approach has been made popular by LIME (for *locally interpretable model-agnostic explanations*) [Ribeiro et al., 2016], where the learning set of the surrogate model is built using black-box access to the model to be explained and a specific sampling technique, and with a popular implementation based on a linear classifier. This approach suffers from the shady contours of the notion of interpretability [Lipton, 2017].

c. *Counterfactual causes:* instead of focusing on how the decision was made, it might be enlightening to question the way things could have turned out differently. This notion is captured by the notion of *counterfactual faithfulness* [Wachter et al., 2017, Doshi-Velez et al., 2017]. In our case, that could result in pointing at a subset of jurors that would have reversed the outcome, would have they approved the candidate.

d. *Focusing on the unexpected:* assuming people only ask for clarification when they are surprised by the outcome, relevant explanations can use a cognitive model of the recipient [Kass and Finin, 1988, Miller, 2018]. In multiple criteria decision aiding, [Labreuche, 2011] instantiates this approach by representing the default reasoning of the user with *anchors*, fictitious facts coalescing the user's expectations.

We note that all these approaches proceed from the principle that the decision rule is true, and adopt a *didactic* stance, where the system teaches the users: accompanying them through the rule (for a.), giving them a simplified overview (for b.), anticipating their goals (for c.) or their expectations (for d.).

A dialectical stance based on reference cases We believe the professorial attitude adopted by the system in the main approaches towards explanation detailed above is not suited to address the challenge of *contestability*. What is being contested is the rule itself, hence considering it as an unquestionable axiom may only lead to a dialog of the deaf between the user and the system. The system needs to adopt a lower profile with respect to its own decision-making in order to enter a contradictory procedure.

Articulating this *dialectical* stance around the notion of *reference cases*, an assignment $\alpha^* : \mathbb{X}^* \to \{\textsc{Good}, \textsc{Bad}\}$ has several advantages:

i) The reference cases represent accumulated knowledge,e.g. a compilation of past decisions that are held of representative of a good adjudication.

ii) The reference cases feed the dialog between the jury and the stakeholders with *axioms*, facts that every party consider true.

iii) The reference cases are a mean to channel the dialog into arguing about the commonly acknowledged facts, on the one hand, and the particulars of the plaintiff, on the other hand. It circumvents the need or the opportunity to refer to the situation of any other candidate—thus contributing to preserve some secrecy and preventing a cascading effect, should the decision concerning a plaintiff be reversed.

iv) The reference cases are binding for the jury: whatever the parameter ω—encoding the approved sets and the set of sufficient coalitions—describing the collective functioning of the committee, the sorting rule NCS_ω should coincide with the reference assignment;

$$\forall x^\star \in \mathbb{X}^\star,\ NCS_\omega(x^\star) = a^\star(x^\star)$$

A large base of reference cases can also serve as a guarantee of *procedural regularity*.

The representation of knowledge based on reference cases needs to be complemented by a representation of reasoning specifically designed to reflect the noncompensatory stance we assume.

When there is some jurisprudence $a^\star : \mathbb{X}^\star \to \{\textsc{Good}, \textsc{Bad}\}$, the assignment of a new candidate x can be *necessary*, in the sense that no other assignment is possible.

Definition 7.3 (Necessary assignment w.r.t. reference cases)**.** *Given a positive instance $a^\star$ of Inv-NCS, an alternative $x \in \mathbb{X}$ is necessarily assigned to a category $C \in \{\textsc{Good}, \textsc{Bad}\}$ with respect to the reference assignment $a^\star$ if $a^\star \cup \{(x, \overline{C})\}$ is a negative instance of Inv-NCS, where $\overline{C}$ denotes the category opposite to C.*

Regarding the situation where a candidate challenges the jury about its own outcome, the objective is to justify the classification of the complaining individual with minimal disclosure of the details of the vote.

We outline different cases:

1. the decision concerning the candidate is *necessary* with respect to the jurisprudence: the sorting of the candidate cannot be otherwise, *as long as a number of other classification outcomes are accepted*;
2. the situation of the candidate is ambivalent w.r.t. the jurisprudence. In this case, where the jurisprudence does not constrain the decision of the jury, we consider two design options:
   - 2a. complementing the jurisprudence with a *default rule*; or
   - 2b. leaving the interpretation up to the jury. Then, it may happen that the status of the coalition of jurors having actually approved the candidate is entailed by the jurisprudence, forming a basis for an explanation.

Example 7.6. (Example 7.3 continued.) 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 Bad . On the contrary, the situation of y is ambivalent, as both assignments $(y, \textsc{Good})$ and $(y, \textsc{Bad})$ can be represented together with α.

Necessary decisions entailed by the jurisprudence. An explanation of the *necessity* of an assignment is intrinsically more complex than that for its *possibility*: one needs to prove that it is not possible to separate all pairs of Good and Bad candidates on at least one point of view. The proof relies on some deadlock that needs to be shown. Formally, this situation manifests itself in the form of an unsatisfiable boolean formula, e.g. given by Corollary 5.7. The unsatisfiability of the entire formula can be reduced to an unsatisfiable subset of clauses (MUS) minimal w.r.t. set inclusion, which are commonly used as certificates of infeasibility, and can also be leveraged to produce *explanations* [Junker, 2004, Besnard et al., 2010, Geist and Peters, 2017]. In the case of the necessary decisions by approval sorting with a reference assignment, any MUS

pinpoints a set of pairs of alternatives in $(\alpha^{-1}(\textsc{Good}) \cup \{x\}) \times \alpha^{-1}(\textsc{Bad})$ that cannot be discriminated simultaneously according to the points of view.

Example 7.7. (Example 7.6 continued.) 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 7.7 can be generalized and formalized into an *argument scheme*, with *premises*: i) a k-tuple of pairs $\langle (g^1,b^1),\ldots,(g^k,b^k)\rangle$ of alternatives with opposite assignment, ii) a subset of points of view $B \subseteq N$ with cardinality $k-1$, such that, according to all points of view $i \notin B$, $b^j \succ_i g^j$ for all j, and, according to all points of view $i \in B$ the intervals $]b^1, g^1]_i, \ldots,]b^k, g^k]_i$ are pairwise disjoint.

Clearly, this explanation technique, inspired by the pigeonhole principle[e], is *sound*: the existence of an argument instantiating the premises of this scheme is a sufficient condition for the infeasibility of representing the given assignment in the noncompensatory model, which in turn yields the *conclusion* that the candidate x is necessarily assigned to the other category.

If we assume that the cognitive burden demanded by an explanation along the lines of this argument scheme increases with the number of its premises, we derive an implicit hierarchy among the necessary decisions supported by the scheme, with a nesting $\mathcal{E}^{ph}_1 \subseteq \mathcal{E}^{ph}_2 \subseteq \cdots \subseteq \mathcal{E}^{ph}_{|N|+1}$, where $\mathcal{E}^{ph}_k$ denotes the set of decisions supported by a pigeonhole-inspired scheme with premises referencing at most k pairs of alternatives with opposite assignment. $\mathcal{E}^{ph}_1$ is exactly the set of decisions stemming from Pareto dominance, where a candidate is either at least as good as a reference alternative in the Good category, or at most as good as a reference alternative in the Bad category.

Is this explanation technique *complete*? The question of deciding if this scheme captures a necessary condition, *i.e.* if any decision entailed by the jurisprudence can be supported by such an explanation, is left open.

Conjecture 7.3.

$$\mathcal{E}^{ph}_{|N|+1} = \{(x,c) \in \mathbb{X} \times \{C^1 \prec \cdots \prec C^p\} : x \textit{ is necessarily assigned to } c\}.$$

[e]If there are strictly more pigeons than pigeonholes, then there is a pigeonhole containing at least two pigeons. Sophisticated people are interested in socks and drawers, and call it Dirichlet's theorem.

Ambivalent situations. It may happen that, for a given candidate, both assignments to Good and to Bad are possible. This situation is obviously all the more frequent as the reference set is small, or the number of points of view is high.

To handle ambivalent cases, a design option would consist in constraining the decision of the committee, either favorably (e.g. following an *innocent unless proven guilty* principle) or unfavorably (e.g. following a *precautionary principle*). In both case, this amounts to completely mechanize the decision procedure, with the advantages of guaranteeing total procedural regularity, and eschewing the need for gathering a real jury, but also incurring the inconvenience of having a completely public procedure that becomes gameable, and depriving the jury of having the possibility to exercise judgment in the adjudication of the ambivalent cases. Concerning the explanation of decisions, having a default rule entails that only the decisions that are necessarily contrary to the rule need to be explained. This option can therefore be seen as a strong enforcement of the accountability requirements: no decision can be taken unless it is duly motivated.

Another, less drastic, option consists in giving the freedom of choice to the committee. In this case, as opposed to the situation where the decision is entailed by the jurisprudence, and where the committee just needs to make obvious the link between the current case and the reference cases, the committee needs to disclose some additional information concerning its inner functioning. In some cases, though, revealing some information concerning the approved sets of the jurors may suffice to explain a specific outcome, thanks to Lemma 5.5.

For an unhappy candidate y assigned to Bad, suppose there exists $B \subset N$ such that:

$$\forall \mathbf{\Large 😞} \in B, \exists x \in \mathbb{X}^\star : \; y \prec_{\mathbf{\Large 😞}} x \text{ and } x \notin \mathcal{A}_{\mathbf{\Large 😞}}; \tag{7.1}$$

and

$$\exists b \in \alpha^{\star -1}(\textsc{Bad}) : \forall \mathbf{\Large 😞} \in N \setminus B, \; b \in \mathcal{A}_{\mathbf{\Large 😞}}. \tag{7.2}$$

The condition (7.3) certifies the candidate is disapproved by all the jurors in B, while the condition (7.4) assesses the remaining jurors do not form a sufficient coalition to qualify the candidate as Good.

Example 7.8. (Example 7.6 continued)

- According to $\mathbf{\Large 😞}^1$, y is disapproved, as it is worse than $e \notin \mathcal{A}_{\mathbf{\Large 😞}^1}$[f];

[f]Note that the explanation could lean on $c \notin \mathcal{A}_{\mathbf{\Large 😞}^1}$, but it might be more convincing to mention, if possible, reference cases that belong to $\alpha^{\star -1}(\textsc{Bad})$, in order to exacerbate the lack of fitness of the candidate—even though, here, c is preferred to e according to $\mathbf{\Large 😞}^1$.

- according to ⚲^3, y is disapproved, as it is worse than $b \notin \mathcal{A}_{\text{⚲}^3}$;
- according to ⚲^5, y is disapproved, as it is worse than $f \notin \mathcal{A}_{\text{⚲}^5}$.
- Furthermore, being approved by ⚲^2 and ⚲^4 is not enough to warrant access to the Good category, as illustrated by e.

Hence, y is assigned to the Bad category.

Conversely, a complaint about a candidate y assigned to Good might be addressed by finding a subset of jurors $B \subset N$ such that:

$$\forall \text{⚲} \in B, \exists x \in \mathbb{X}^\star : \; y \succ_{\text{⚲}} x \text{ and } x \in \mathcal{A}_{\text{⚲}}; \tag{7.3}$$

and

$$\exists g \in \alpha^{\star -1}(\textsc{Good}) : \forall \text{⚲} \in N \setminus B, \; g \notin \mathcal{A}_{\text{⚲}}. \tag{7.4}$$

The condition (7.3) certifies the candidate is approved by all the jurors in B, while the condition (7.4) assesses they form a sufficient coalition to qualify the candidate as Good.

These explanations do not cover every possible configuration. Indeed, they do not reference the actual, latent, set of sufficient coalitions of jurors but only its bounds, given the approved sets, established by Lemma 5.5—the upper bound $\mathcal{P}(N) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha^\star)$ when the contestation concerns a candidate deemed Bad, the lower bound $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha^\star)$ when the contestation concerns a candidate deemed Good. Figure 7.1 illustrates these bounds of the version space of the sufficient coalitions of jurors, given the approved sets, corresponding to the case detailed in Example 7.8. In order to ensure every decision of the jury can be supported by an explanation, the remaining cases, corresponding to candidates approved by a coalition of jurors that is neither necessarily nor impossibly sufficient given the actual approved sets, could be handled by a default rule.

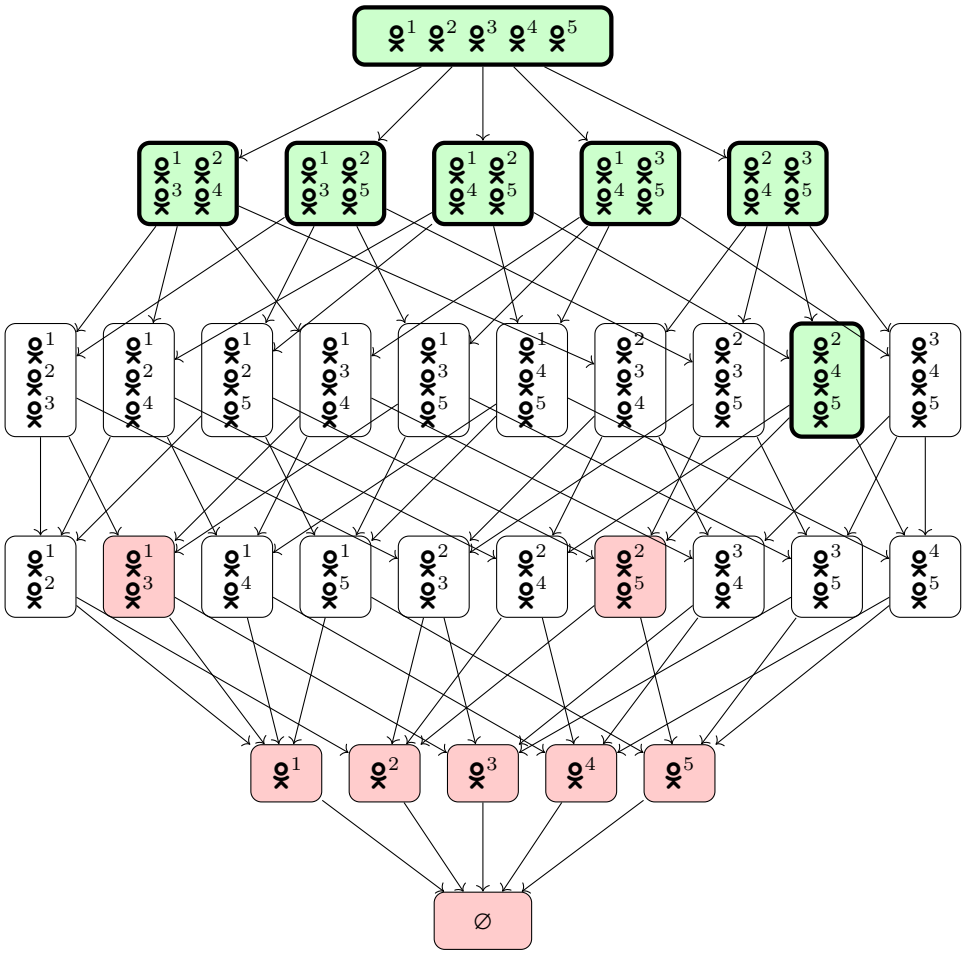

Figure 7.1: Version space of the sufficient coalitions of a noncompensatory sorting model, with given approved sets, corresponding to Example 7.8. Necessarily sufficient coalitions are in green with thick borders, necessarily insufficient coalitions are in red and ambivalent coalitions—possibly sufficient or insufficient—are in white.

Conclusion

8

Conclusion

8.1 Summary of our contributions

We summarize below the main results obtained in this PhD thesis, following the structure of the document.

Contributions to the problem of comparing with an additive model.

a. We described the aggregation procedure yielded by the (robust) elicitation of an additive value model in a KR framework, with the preference information as a knowledge base and adjudication as inference (mostly published in [Belahcene et al., 2017a]);

b. we provided knobs permitting to account for variants of the additive value model, either more constrained, or allowing for a richer language (unpublished);

c. we proposed a framework permitting to compute and represent the entire necessary preference relation (mostly published in [Belahcene et al., 2017a]);

d. we proposed an explanation engine for this model, based on cancellation, permitting to tie conclusion drawn to premises that are PI statements; various representations are discussed, completeness and soundness are established, computation of simple explanations is NP complete (unpublished);

e. we proposed another explanation engine for the additive value model, based on a *divide and conquer* principle consisting in breaking down

a complex statement into simpler statements combined by transitivity; the scheme is sound, but there is a trade-off between simplicity and completeness—we furthermore derive excellent completeness, simplicity and computational results when considering a classical restriction on the preference information. (published in [Belahcene et al., 2017a])

Contributions to the problem of noncompensatory sorting.

a. We described in a unified framework some of the variants of the noncompensatory sorting model (unpublished);

b. we proposed a formulation based on Boolean satisfiability, representing the inverse-NCS problem in a sound and complete manner that explicitly represents the parameter space, and thus requires an exponential number of variables and clauses w.r.t. the number of points of view (published in [Belahcene et al., 2018c]);

c. we experimentally assessed the computational relevance of this formulation, by comparing it to the state-of-the-art elicitation procedure for MR-Sort, a noncompensatory sorting model where sufficient coalitions of criteria are normatively assumed to be compactly represented by a majority rule; results tend to show the proposed formulation is approximately 50 times faster than the baseline, while being moderately more prone to overfitting (published in [Belahcene et al., 2018c]);

d. we proposed and proved a characterization of the positive instances eschewing the explicit representation of the power set of points of view, and proposed a second formulation based on Boolean satisfiability, representing the inverse-NCS problem in a sound and complete manner, with a polynomial number of variables and clauses (published in [Belahcene et al., 2018a] in the case of two categories);

e. we (briefly) discussed the potential contribution of an approach based on Inv-NCS to the robust elicitation of the model, towards an accountable decision aiding framework (unpublished);

f. we envisioned the potential contribution of Inv-NCS to the elaboration of an accountable-by-design sorting procedure based on the approval of candidates by jurors; complete and sound certificates for procedural regularity are provided in the form of argument schemes, even though simplicity cannot guaranteed in general; moreover, sound certificates addressing potential contest are proposed, in support of necessary assignments, also in the form of argument schemes (published in [Belahcene et al., 2018a]).

8.2 Open questions and work in progress

There are a number of promising future work which can be undertaken very soon (or are already in progress, as we shall see). We begin with those pertaining to Part I, hence assuming an additive value model:

1. Compare the two approaches to the explanation of necessary pairwise preference, assuming an additive value model—one based on cancellation, the other based on simple transitive sequences. By comparison here we mean both a deeper theoretical understanding of their relative properties (as length of sequences etc.); but also experiments involving human users, so as to evaluate the cognitive burden and human understandability of the proposed explanations [Nunes et al., 2014, Miller, 2018];

2. An obvious open question regards the possibility (or not) to avoid the repetition of the conclusion, in the cancellative explanations of necessary pairwise preference, assuming an additive value model. We conjecture that it is impossible to avoid, but this remains to be proven;

3. Develop the extensions of the additive model towards cardinality. There are several opportunities that we may think of: rounding of the covector coefficients, explanation of residual regret, preference information relative to the intensity of preference, explanation *of* or *with* intensity of preferences. A first interesting step could be to account for quaternary preferences, thus allowing statements of the form *"a is more preferred to b than c is preferred to d"* [Bana e Costa and Vansnick, 1995, Bouyssou and Pirlot, 2004].

Regarding the Noncompensatory Sorting model studied in Part II, we start by pinpointing a couple of questions already mentioned in the document, and then enumerate direct variants of the problem which could be considered:

1. We left open an intriguing but potentially challenging question regarding the completeness of the explanation scheme based on the pigeonhole principle, for necessary adjudications of noncompensatory sorting. If we were to answer this conjecture in the positive, this would illustrate how pervasive this principle could be and potentially offer insights for other problems. We note that this would also readily imply an upper bound on the complexity of explanations (taken as the arity).

2. While we provide first results regarding lower bounds on the length of explanations (for instance regarding the number of viewpoints involved), we could also exploit *communication complexity* [Kushilevitz and Nisan, 1997] notions to come up with general lower bounds results for the

problem under consideration, e.g. the number of bits that need to be exchanged in order to solve the Inv-NCS decision problem.

3. As we came up with different SAT formulations for Inv-NCS, it certainly makes sense to experimentally assess their computational relevance, and compare in particular the compact to the non-compact one already provided (joint work with Oumeima Khaled and Ali Tilil, to appear in [Belahcene et al., 2018b]—presented in Appendix C).

Finally, there are straightforward variants of the model investigated in this document:

1. For the case of two categories, we came up with a full battery of results (characterization results for the cases where approved sets are know or not, and subsequent compact SAT formulation). The same exercise should be done for the model allowing to sort into *more than two categories*. We do not foresee any specific difficulty here, and believe similar results should be obtained in the near future.

2. There are other very interesting noncompensatory models existing in the literature. Can we extend the principle of representing inverse problem for disaggregation to these models? In joint work with Marc Pirlot and Olivier Sobrie, to appear in [Belahcene et al., 2018d]—presented in Appendix D— we take up this research agenda and investigate the model of *ranking with multiple reference points* [Rolland, 2013].

3. Another idea could be to bridge the noncompensatory sorting models (which are based on a limitation *a priori* on the number of ordinal categories, and constrains as a consequence the language allowed to the decision maker with a similar cap for all points of view), to L^1-optimization techniques ('lasso') that yield the minimal number of total categories [Sokolovska et al., 2017].

4. Finally, a natural feature required in many applications is to allow to express additional constraints, e.g. cardinality constraints on the categories (*"We would like to have at least 5 acceptable projects to fund"*). This can be done inside the Boolean satisfiability framework, using SMT (SAT modulo theory).

8.3 Perspectives

Considering a not-so-perfect preference information. When considering to put our contributions to the test consisting in supporting a real-world decision aiding process with explanation, we foresee that the principal ob-

stacle will be the assumption of having a perfect and consistent preference information.

Therefore, the challenge of finding a principled way to restore this consistency (or deal with inconsistency) in an accountable manner, needs to be addressed. Several promising approaches have been proposed on the basis that some assumptions—either preference statements or decision theoretic properties—might simply be *wrong*, and should be discarded:

- considering maximally consistent subsets of statements [Mousseau et al., 2003];
- relaxing the aggregation model—assuming models form a lattice, seen as a partially nested structure—until a model sufficiently expressive to accommodate for the preference information is found [Ouerdane, 2010, Greco et al., 2014];
- using a numerical estimation of inconsistency, such as a *belief function* [Destercke, 2018].

An intriguing question concerns the representation of knowledge about the aggregation procedure representing the views of the decision maker. While we have rejected the notion of learning a compiled representation—a ball—in favor of a list of statements—a polytope, when the model is linear—it seems reasonable to assume that, given sufficient time, the elicitation process might eventually converge towards a single value of the preference parameter. When does it become sensible to move from one representation to another? Studying this *phase transition* presumably requires a non-dialectic framework that tolerates noisy inputs.

Languages. To represent the way of reasoning about preferences embodied by the additive value model and the noncompensatory sorting model, we have considered several paradigms:

- Linear programming (LP), with continuous variables;
- Boolean satisfiability (SAT), with binary variables;
- Mixed integer programming (MIP), which is an extension of LP that allows for integer variables, but also extends SAT;

LP is polynomial, while SAT and MIP are NP-complete. Structurally, MIP is far more expressive than SAT and LP, but is practically much slower, and does not propose handy, generic certificates like LP or SAT. Replacing MIP by SAT, in the case of Inv-NCS problem, permitted huge gains in terms of computation time. Many more languages, of intermediate complexity and expressiveness, can be considered:

- MaxSAT, where a SAT formulation is complemented with an objective function, so that the number of satisfied clauses is maximal, allows to 'best satisfy' an unsatisfiable instance;
- pw-MaxSAT, allows to weigh ('weighted') or omit ('partial') the contribution of clauses to the objective function;
- *pseudo Boolean functions* (pBf) and *answer set programming* (ASP) offer an even richer language—yet less so than MIP—that allow to natively represent e.g. cardinality constraints [Berre et al., 2018].

Dropping the barrier between solver and solved. In this work, we have repeatedly leveraged the existence of powerful solvers for MIP and SAT formulations. This black-box approach allowed us to eschew the task of developing a dedicated solver for our problems, while benefiting of the latest advances. As our approach is based on certificates of infeasibility, it might be beneficial to:

1. have a look at the representation of this type of certificates inside the various solvers, both as a source for new ideas for explanation, and as a way to avoid the reconstruction of such a certificate outside of the solver, involving a large number of black-bow queries;
2. design solvers tailored to our problems and our needs for certificates.

This is the topic of the current CNRS PEPS project *SAT4Ex*, involving the CRIL laboratory and ourselves.

Models. Throughout this work, the selection of an aggregation model has been normative, and considered as belonging to the context. Models embody assumptions about the decision-theoretic stance of the decision maker w.r.t. a specific decision situation; therefore, a given model may prove inappropriate to a given situation. For instance, the additive value model has two well-known blind spots:

- *incomparability*—the preference structure described by any value model is a total preorder. Given two alternatives, only three outcomes are possible: strict preference for the former, strict preference for the latter, or indifference. It has been shown (e.g. [Deparis et al., 2012, Dubois et al., 2008]) that decision makers often spontaneously offer a fourth type of answer: incomparability. This second type of symmetric relation has an epistemic meaning that differs from the one expressed by indifference——in one case, the subject expresses that they know the alternatives compares equally, while in the other they express that they have no knowledge.

This epistemic difference may translate into different properties, and therefore may need to be accounted for by the preference model.

- *interactions between the influence of the points of view*—first-order cancellation can be interpreted as a statement of preferential independence. It offers the opportunity to reason ceteris paribus, i.e. everything else being equal, without any consideration for the actual values of the things 'being equal'. This fundamental feature of the additive model can, and has often been, defeated in practical applications. We recall some famous refutations of this property, e.g. Fishburn's iterated menus, or values of bundles [Fisher, 1892, Bouveret et al., 2016], with substitution or complementarity effects between objects.

  Many models account for interactions between the influence of the points of view, such as Generalized additive models (GAI) [Fishburn, 1967b], or Choquet integral [Grabisch and Labreuche, 2010]. These models, however, are practically defined as corrections to the baseline offered by the additive value model.

Another issue is the *informational complexity* of models: how much preference information is needed to elicit 'sufficiently'. It seems crucial to select a model of a complexity proportionate to the time budget allocated to the decision process. This assumption can be relaxed in at least two directions:

- obviously, considering other aggregation models than those considered here;
- considering the model selection as defeasible [Ouerdane, 2010].

Planning. Considering preference elicitation as a planning problem is not a new idea (for instance, Boutilier [Boutilier, 2002] proposes a POMDP formulation). Similarly, approaching argument generation as a planning problem is natural, and not completely new either [Cawsey, 1993]. 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. 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), as well as –perhaps more prospectively– interleave the elicitation and explanation process.

Applications. Finally, a very valuable achievement would be to implement a proof of concept:

- addressing a real-world decision situation, and proposing to assist an analyst during the elicitation of an additive value model, or of a noncompensatory sorting model;
- participating in the elaboration of an *accountable-by-design* procedure, maybe for an *administrative algorithm* [Cozic and Valarcher, 2017], or any other agent required to account for the decision they take [Kroll et al., 2017].

Bibliography

[Aingworth et al., 1996] Aingworth, D., Chekuri, C., and Motwani, R. (1996). Fast estimation of diameter and shortest paths (without matrix multiplication). In *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '96, pages 547–553.

[Alvarez, 2004] Alvarez, I. (2004). Explaining the result of a Decision Tree to the End-User. In *Proceedings of the 16th European Conference on Artificial intelligence (ECAI'16)*, pages 411–415. IOS Press.

[Amor et al., 2015] Amor, N. B., Dubois, D., Gouider, H., and Prade, H. (2015). Possibilistic conditional preference networks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 36–46. Springer.

[Amor et al., 2016] Amor, N. B., Dubois, D., Gouider, H., and Prade, H. (2016). Graphical models for preference representation: an overview. In *International Conference on Scalable Uncertainty Management*, pages 96–111. Springer.

[Angilella et al., 2010] Angilella, S., Greco, S., and Matarazzo, B. (2010). Non-additive robust ordinal regression: A multiple criteria decision model based on the choquet integral. *European Journal of Operational Research*, 201(1):277–288.

[Arrow, 1950] Arrow, K. J. (1950). A difficulty in the concept of social welfare. *The Journal of Political Economy*, pages 328–346.

[Bana e Costa and Vansnick, 1995] Bana e Costa, C. and Vansnick, J. (1995). General overview of the MACBETH approach. In Pardalos, P., Siskos, Y., and Zopounidis, C., editors, *Advances in Multicriteria Analysis*, pages 93–100. Kluwer Academic Publishers, Dordrecht.

[Bana e Costa et al., 2008] Bana e Costa, C. A., Lourencco, J. C., Chagas, M. P., and Bana e Costa, J. C. (2008). Development of reusable bid evaluation models for the portuguese electric transmission company. *Decision Analysis*, 5(1):22–42.

[Bana e Costa and Vansnick, 1994] Bana e Costa, C. A. and Vansnick, J.-C. (1994). MACBETH–an interactive path towards the construction of cardinal value functions. *International transactions in operational Research*, 1(4):489–500.

[Belahcene et al., 2018a] Belahcene, K., Chevaleyre, Y., Labreuche, C., Maudet, N., Mousseau, V., and Ouerdane, W. (2018a). Accountable approval sorting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 70–76. International Joint Conferences on Artificial Intelligence Organization.

[Belahcene et al., 2018b] Belahcene, K., Khaled, O., Mousseau, V., Ouerdane, W., and Tlili, A. (2018b). A new efficient SAT formulation for learning NCS models: numerical results. In *4th international workshop from Multiple Criteria Decision Aid to Preference Learning (DA2PL'18)*.

[Belahcene et al., 2017a] Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., and Ouerdane, W. (2017a). Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183.

[Belahcene et al., 2017b] Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., and Ouerdane, W. (2017b). A model for accountable ordinal sorting. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 814–820.

[Belahcene et al., 2018c] Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., and Ouerdane, W. (2018c). An efficient sat formulation for learning multiple criteria non-compensatory sorting rules from examples. *Computers & Operations Research*, 97:58 – 71.

[Belahcene et al., 2018d] Belahcene, K., Mousseau, V., Pirlot, M., Ouerdane, W., and Sobrie, O. (2018d). Ranking with Multiple Points: Efficient Elicitation and Learning Procedures. In *4th international wokshop from Multiple Criteria Decision Aid to Preference Learning (DA2PL'18)*.

[Ben-Tal et al., 2009] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press.

[Benabbou et al., 2016] Benabbou, N., Perny, P., and Viappiani, P. (2016). A regret-based preference elicitation approach for sorting with multicriteria

reference profiles. In *From Multicriteria Decision Making to Preference Learning (DA2PL'16)*, Paderborn, Germany.

[Benabbou et al., 2017] Benabbou, N., Perny, P., and Viappiani, P. (2017). Incremental elicitation of choquet capacities for multicriteria choice, ranking and sorting problems. *Artificial Intelligence*, 246:152–180.

[Berre et al., 2018] Berre, D. L., Marquis, P., Mengel, S., and Wallon, R. (2018). Pseudo-boolean constraints from a knowledge representation perspective. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1891–1897. International Joint Conferences on Artificial Intelligence Organization.

[Besnard et al., 2010] Besnard, P., Grégoire, É., Piette, C., and Raddaoui, B. (2010). MUS-based generation of arguments and counter-arguments. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pages 239–244.

[Besse et al., 2017] Besse, P., Castets-Renard, C., and Garivier, A. (2017). Loyalty of algorithmic decisions. working paper or preprint.

[Bisdorff et al., 2015] Bisdorff, R., Dias, L. C., Meyer, P., Mousseau, V., and Pirlot, M. (2015). *Evaluation and decision models with multiple criteria: Case studies*. Springer.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[Bounhas et al., 2018] Bounhas, M., Pirlot, M., and Prade, H. (2018). Predicting preferences by means of analogical proportions. In *Case-Based Reasoning Research and Development - 26th International Conference, ICCBR 2018, Proceedings*, pages 515–531.

[Bous et al., 2010] Bous, G., Fortemps, P., Glineur, F., and Pirlot, M. (2010). Acuta: A novel method for eliciting additive value functions on the basis of holistic preference statements. *EJOR*, 206(2):435–444.

[Boutilier, 2002] Boutilier, C. (2002). A POMDP formulation of preference elicitation problems. In *Eighteenth National Conference on Artificial Intelligence*, pages 239–246, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Boutilier et al., 2004] Boutilier, C., Brafman, R. I., Domshlak, C., Hoos, H. H., and Poole, D. (2004). Cp-nets: A tool for representing and reasoning withconditional ceteris paribus preference statements. *Journal of artificial intelligence research*, 21:135–191.

[Boutilier et al., 2006] Boutilier, C., Patrascu, R., Poupart, P., and Schuurmans, D. (2006). Constraint-based optimization and utility elicitation using the minimax decision criterion. *Artificial Intelligence*, 170(8-9):686–713.

[Boutilier et al., 2010] Boutilier, C., Regan, K., and Viappiani, P. (2010). Simultaneous elicitation of preference features and utility. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI'10*, Atlanta, Georgia, USA.

[Bouveret et al., 2016] Bouveret, S., Chevaleyre, Y., and Maudet, N. (2016). Fair allocation of indivisible goods. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors, *Handbook of Computational Social Choice*, pages 284–310. Cambridge University Press.

[Bouyssou, 1986] Bouyssou, D. (1986). Some remarks on the notion of compensation in mcdm. *EJOR*, 26(1):150–160.

[Bouyssou and Marchant, 2007a] Bouyssou, D. and Marchant, T. (2007a). An axiomatic approach to noncompensatory sorting methods in mcdm, i: The case of two categories. *EJOR*, 178(1):217–245.

[Bouyssou and Marchant, 2007b] Bouyssou, D. and Marchant, T. (2007b). An axiomatic approach to noncompensatory sorting methods in mcdm, ii: More than two categories. *EJOR*, 178(1):246–276.

[Bouyssou et al., 2000] Bouyssou, D., Marchant, T., Perny, P., Pirlot, M., Tsoukiàs, A., and Vincke, P. (2000). *Evaluation and decision models: a critical perspective*, volume 32 of *International Series in Operations Research and Management Science*. Kluwer Academic Publishers.

[Bouyssou et al., 2006] Bouyssou, D., Marchant, T., Pirlot, M., Tsoukiàs, A., and Vincke, P. (2006). *Evaluation and decision models with multiple criteria: Stepping stones for the analyst*. Springer Verlag.

[Bouyssou and Pirlot, 2004] Bouyssou, D. and Pirlot, M. (2004). 'additive difference'models without additivity and subtractivity. *Journal of Mathematical Psychology*, 48(4):263–291.

[Brabant et al., 2018] Brabant, Q., Couceiro, M., Dubois, D., Prade, H., and Rico, A. (2018). Extracting Decision Rules from Qualitative Data via Sugeno Utility Functionals. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations - 17th International Conference, IPMU 2018, Proceedings, Part I*, pages 253–265.

[Brandt et al., 2016] Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (2016). *Handbook of Computational Social Choice*. Cambridge University Press, New York, NY, USA, 1st edition.

[Braziunas and Boutilier, 2007] Braziunas, D. and Boutilier, C. (2007). Minimax regret based elicitation of generalized additive utilities. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI'07)*, pages 25–32, Virginia, United States. AUAI Press.

[Buchanan and Shortliffe, 1984] Buchanan, B. G. and Shortliffe, E. H. (1984). *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Boston, MA, USA.

[Cailloux and Endriss, 2014] Cailloux, O. and Endriss, U. (2014). Eliciting a suitable voting rule via examples. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI'14)* , volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 183–188. IOS Press.

[Cailloux and Endriss, 2016] Cailloux, O. and Endriss, U. (2016). Arguing about voting rules. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'16)*, pages 287–295. International Foundation for Autonomous Agents and Multiagent Systems.

[Cailloux and Meinard, 2018] Cailloux, O. and Meinard, Y. (2018). A formal framework for deliberated judgment. *arXiv preprint arXiv:1801.05644*.

[Cawsey, 1993] Cawsey, A. (1993). Planning interactive explanations. *International Journal of Man-Machine Studies*, 38(2):169 – 199.

[Charnetski and Soland, 1978] Charnetski, J. R. and Soland, R. M. (1978). Multiple-attribute decision making with partial information: The comparative hypervolume criterion. *Naval Research Logistics Quarterly*, 25(2):279–288.

[Colorni and Tsoukiàs, 2013] Colorni, A. and Tsoukiàs, A. (2013). What is a decision problem? preliminary statements. In Perny, P., Pirlot, M., and Tsoukiàs, A., editors, *Algorithmic Decision Theory*, pages 139–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Condorcet, 1785] Condorcet (1785). *Essai sur l'application de l'analyse a la probabilité des decisions rendues a la pluralité des voix [microform] / par M. le Marquis de Condorcet*. Imprimerie royale Paris.

[Cook, 1971] Cook, S. A. (1971). The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pages 151–158, New York, USA. ACM.

[Cormen et al., 2001] Cormen, T. H., Stein, C., Rivest, R. L., and Leiserson, C. E. (2001). *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition.

[Corrente et al., 2014] Corrente, S., Greco, S., Kadziński, M., and Słowiński, R. (2014). *Robust Ordinal Regression*, pages 1–10. American Cancer Society.

[Cozic and Valarcher, 2017] Cozic, M. and Valarcher, P. (2017). The design of Administrative Algorithms. the ALGOCIT team. Presented at the Social Responsibility of Algorithms - SRA 2017 interdisciplinary workshop. Paris, Dauphine.

[Debreu, 1960] Debreu, G. (1960). Topological methods un cardinal utility theory. *Mathematical Methods in the Social Sciences*.

[Deparis et al., 2012] Deparis, S., Mousseau, V., Öztürk, M., Pallier, C., and Huron, C. (2012). When conflict induces the expression of incomplete preferences. *European Journal of Operational Research*, 221(3):593–602.

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

[Devaud et al., 1980] Devaud, J., Groussaud, G., and Jacquet-Lagreze, E. (1980). UTADIS: Une methode de construction de fonctions d'utilite additives rendant compte de jugements globaux. In *European working group on MCDA, Bochum , Germany*.

[Dickerson et al., 2014] Dickerson, J. P., Goldman, J. R., Karp, J., Procaccia, A. D., and Sandholm, T. (2014). The computational rise and fall of fairness. In *AAAI*, volume 14, pages 1405–1411.

[Doshi-Velez et al., 2017] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., and Wood, A. (2017). Accountability of AI under the law: The role of explanation. *CoRR*, abs/1711.01134.

[Dubois et al., 2008] Dubois, D., Fargier, H., and Bonnefon, J.-F. (2008). On the qualitative comparison of decisions having positive and negative features. *Journal of Artificial Intelligence Research*, 32:385–417.

[Dubois et al., 2005] Dubois, D., Kaci, S., and Prade, H. (2005). Expressing preferences from generic rules and examples - A possibilistic approach without aggregation function. In Godo, L., editor, *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2005)*, pages 293–304. Springer-Verlag.

[Dubois and Prade, 1988] Dubois, D. and Prade, H. (1988). *Possibility theory*. Plenum Press, New-York.

[Dubois and Prade, 2008] Dubois, D. and Prade, H. (2008). An introduction to bipolar representations of information and preference. *International Journal of Intelligent Systems*, 23(8):866–877.

[Dung, 1995] Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person games. *Artificial Intelligence*, 77(2):321–358.

[Even and Tarjan, 1975] Even, S. and Tarjan, R. E. (1975). Network flow and testing graph connectivity. *SIAM J. Comput.*, 4(4):507–518.

[Figueira et al., 2005] Figueira, J., Mousseau, V., and Roy, B. (2005). Electre methods. In *Multiple criteria decision analysis: State of the art surveys*, pages 133–153. Springer.

[Figueira et al., 2009] Figueira, J. R., Greco, S., and Slowinski, R. (2009). Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. *European Journal of Operational Research*, 195(2):460–486.

[Fishburn, 1967a] Fishburn, P. C. (1967a). Additive utilities with finite sets: Applications in the management sciences. *Naval Research Logistics Quarterly*, 14(1):1–13.

[Fishburn, 1967b] Fishburn, P. C. (1967b). Interdependence and additivity in multivariate, unidimensional expected utility theory. *International Economic Review*, 8(3):335–342.

[Fishburn, 1976] Fishburn, P. C. (1976). Noncompensatory preferences. *Synthese*, 33(2/4):393–403.

[Fishburn, 1997] Fishburn, P. C. (1997). Failure of cancellation conditions for additive linear orders. *Journal of Combinatorial Designs*, 5(5):353–365.

[Fisher, 1892] Fisher, I. (1892). *Mathematical investigations in the theory of value and prices, and appreciation and interest.*

[Fürnkranz and Hüllermeier, 2010] Fürnkranz, J. and Hüllermeier, E. (2010). *Preference Learning*. Springer-Verlag, Berlin, Heidelberg, 1st edition.

[Gammerman et al., 1998a] Gammerman, A., Vovk, V., and Vapnik, V. (1998a). Learning by transduction. In *In Uncertainty in Artificial Intelligence*, pages 148–155. Morgan Kaufmann.

[Gammerman et al., 1998b] Gammerman, A., Vovk, V., and Vapnik, V. (1998b). Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 148–155, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Geffner, 2018] Geffner, H. (2018). Model-free, model-based, and general intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI'18*, pages 10–17, Stockholm, Sweden.

[Geist and Peters, 2017] Geist, C. and Peters, D. (2017). Computer-aided methods for social choice theory. In Endriss, U., editor, *Trends in Computational Social Choice*, chapter 13, pages 249–267. AI Access.

[Giarlotta and Greco, 2013] Giarlotta, A. and Greco, S. (2013). Necessary and possible preference structures. *Journal of Mathematical Economics*, 49(2):163 – 172.

[Goldstein, 1991] Goldstein, W. M. (1991). Decomposable threshold models. *Journal of Mathematical Psychology*, 35(1):64 – 79.

[Gonzales, 2003] Gonzales, C. (2003). Additive Utility Without Restricted Solvability on Every Component. *Journal of Mathematical Psychology*, 47(1):47–65.

[Gonzales and Perny, 2004] Gonzales, C. and Perny, P. (2004). Gai networks for utility elicitation. In *Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning*, pages 224–233. AAAI Press.

[Goodman and Flaxman, 2017] Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

[Grabisch and Labreuche, 2010] Grabisch, M. and Labreuche, C. (2010). A decade of application of the choquet and sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175(1):247–286.

[Gérard et al., 2007] Gérard, R., Kaci, S., and Prade, H. (2007). Ranking Alternatives on the Basis of Generic Constraints and Examples - A Possibilistic Approach. In *International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 393–398. International Joint Conference on Artificial Intelligence (IJCAI).

[Greco et al., 2001a] Greco, S., Matarazzo, B., and Slowinski, R. (2001a). Conjoint measurement and rough set approach for multicriteria sorting problems in presence of ordinal criteria. In Colorni, A., Paruccini, M., and Roy, B., editors, *A-MCD-A*, Scientific and technical research series, pages 117–144. Office of Official Publications of the European Communities.

[Greco et al., 2001b] Greco, S., Matarazzo, B., and Slowinski, R. (2001b). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129(1):1–47.

[Greco et al., 2008] Greco, S., Mousseau, V., and Słowiński, R. (2008). Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *EJOR*, 191(2):416–436.

[Greco et al., 2010a] Greco, S., Mousseau, V., and Slowinski, R. (2010a). Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research*, 207(3):1455–1470.

[Greco et al., 2014] Greco, S., Mousseau, V., and Slowinski, R. (2014). Robust ordinal regression for value functions handling interacting criteria. *European Journal of Operational Research*, 239(3):711–730.

[Greco et al., 2010b] Greco, S., Słowinski, R., Figueira, J., and Mousseau, V. (2010b). Robust ordinal regression. In *Trends in Multiple Criteria Decision Analysis*, pages 241–284. Springer Verlag.

[Guitouni and Martel, 1998] Guitouni, A. and Martel, J. (1998). Tentative guidelines to help choosing an appropriate MCDA method. *European Journal of Operational Research*, 109(2):501–521.

[Gunning, 2017] Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). Accessed on 2018-10-08 (http://www.darpa.mil/program/explainable-artificial-intelligence).

[Hammond et al., 1998] Hammond, J., Keeney, R., and Raiffa, H. (1998). Even Swaps: a rational method for making trade-offs. *Harvard Business Review*, march-june:137–149.

[Hazen, 1986] Hazen, G. B. (1986). Partial information, dominance, and potential optimality in multiattribute utility theory. *Operations Research*, 34(2):296–310.

[Hyafil and Boutilier, 2006] Hyafil, N. and Boutilier, C. (2006). Regret-based incremental partial revelation mechanisms. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, pages 672–678, Boston, Massachusetts, USA.

[Jacquet-Lagrèze and Siskos, 1982] Jacquet-Lagrèze, E. and Siskos, Y. (1982). Assessing a set of additive utility functions for multicriteria decision making: the UTA method. *European Journal of Operational Research*, 10:151–164.

[Jacquet-Lagrèze and Siskos, 2001] Jacquet-Lagrèze, E. and Siskos, Y. (2001). Preference disaggregation: 20 years of MCDA experience. *European Journal of Operational Research*, 130(2):233–245.

[Junker, 2004] Junker, U. (2004). Quickxplain: Preferred explanations and relaxations for over-constrained problems. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 167–172, Menlo Park, California. AAAI Press / The MIT Press.

[Kaci, 2011] Kaci, S. (2011). *Working with Preferences: Less Is More*. Cognitive Technologies. Springer.

[Kadzinski et al., 2012] Kadzinski, M., Greco, S., and Slowinski, R. (2012). Selection of a representative value function in robust multiple criteria ranking and choice. *European Journal of Operational Research*, 217(3):541 – 553.

[Karmarkar, 1984] Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311. ACM.

[Karp, 1972] Karp, R. M. (1972). Reducibility among combinatorial problems. In Miller, R. E., Thatcher, J. W., and Bohlinger, J. D., editors, *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations*, pages 85–103. Springer US, Boston, MA.

[Kass and Finin, 1988] Kass, R. and Finin, T. (1988). The need for user models in generating expert system explanation. *Int. J. Expert Syst.*, 1(4):345–375.

[Kazman et al., 2000] Kazman, R., Klein, M., and Clements, P. (2000). *ATAM: Method for Architecture Evaluation*. TECHNICAL REPORT, CMU/SEI-2000-TR-004, http://www.sei.cmu.edu/reports/00tr004.pdf.

[Keeney and Raiffa, 1976] Keeney, R. and Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. J. Wiley, New York.

[Krantz et al., 1971] Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971). *Foundations of measurement*, volume 1: Additive and Polynomial Representations. Academic Press.

[Kroll et al., 2017] Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165.

[Kushilevitz and Nisan, 1997] Kushilevitz, E. and Nisan, N. (1997). Communication complexity. In *Advances in Computers*.

[Labreuche, 2011] Labreuche, C. (2011). A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence Journal*, 175:1410–1448.

[Lahdelma et al., 1998] Lahdelma, R., Hokkanen, J., and Salminen, P. (1998). SMAA - stochastic multiobjective acceptability analysis. *European Journal of Operational Research*, 106(1):137–143.

[Laslier and Sanver, 2010] Laslier, J.-F. and Sanver, M. R. (2010). *Handbook on Approval Voting*. Studies in Choice and Welfare. Springer, Boston.

[Leroy et al., 2011] Leroy, A., Mousseau, V., and Pirlot, M. (2011). Learning the parameters of a multiple criteria sorting method. In Brafman, R., Roberts, F., and Tsoukiàs, A., editors, *Algorithmic Decision Theory*, volume 6992 of *Lecture Notes in Artificial Intelligence*, pages 219–233.

[Lipton, 2017] Lipton, Z. C. (2017). The mythos of model interpretability. *CoRR*, abs/1606.03490.

[Luce and Tukey, 1964] Luce, R. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1):1 – 27.

[Marichal, 2000] Marichal, J.-L. (2000). On sugeno integral as an aggregation function. *Fuzzy Sets and Systems*, 114(3):347 – 365.

[Miller, 2018] Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269.

[Mitchell, 1982] Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, 18(2):203–226.

[Mousseau, 2003] Mousseau, V. (2003). *Elicitation des préférences pour l'aide multicritère à la décision*. PhD thesis, Mémoire présenté en vue de l'obtention de l'habilitation à diriger des recherches, Université Paris-Dauphine.

[Mousseau et al., 2003] Mousseau, V., Dias, L., Figueira, J., Gomes, C., and Clímaco, J. (2003). Resolving inconsistencies among constraints on the parameters of an MCDA model. *European Journal of Operational Research*, 147(1):72–93.

[Mousseau and Slowinski, 1998] Mousseau, V. and Slowinski, R. (1998). Inferring an ELECTRE TRI model from assignment examples. *Journal of Global Optimization*, 12(2):157–174.

[Mousseau et al., 2000] Mousseau, V., Słowiński, R., and Zielniewicz, P. (2000). A user-oriented implementation of the electre-tri method integrating preference elicitation support. *Computers & operations research*, 27(7):757–777.

[Mustakoji and Hamalainen, 2005] Mustakoji, J. and Hamalainen, R. (2005). A preference programming approach to make the even swaps method even easier. *Decision Analysis*, 2(2):110–123.

[Mustakoji and Hamalainen, 2007] Mustakoji, J. and Hamalainen, R. (2007). Smart-swaps: a decision support system for multicriteria decision analysis with the even swaps method. *Decision Support Systems*, 44(1):313–325.

[Nunes et al., 2014] Nunes, I., Miles, S., Luck, M., Barbosa, S., and Lucena, C. (2014). Pattern-based explanation for automated decisions. In *Proceedings of the 21st European Conference on Artificial intelligence*, pages 669–674. IOS Press.

[Ok, 2002] Ok, E. A. (2002). Utility representation of an incomplete preference relation. *Journal of Economic Theory*, 104(2):429 – 449.

[Ouerdane, 2009] Ouerdane, W. (2009). *Multiple Criteria Decision Aiding: a Dialectical Perspective*. PhD thesis, University Paris-Dauphine, Paris,France.

[Ouerdane, 2010] Ouerdane, W. (2010). Mulitple criteria decision aiding: a dialectical perspective. *4OR: A Quarterly Journal of Operations Research*, 9(4):429–432.

[Peleg, 2002] Peleg, B. (2002). Chapter 8 game-theoretic analysis of voting in committees. In *Handbook of Social Choice and Welfare*, volume 1 of *Handbook of Social Choice and Welfare*, pages 395 – 423. Elsevier.

[Perny, 2000] Perny, P. (2000). *Modélisation des préférences, agrégation multi-critère 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.

[Pettit, 1997] Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government*. Oxford University Press.

[Pettit, 2000] Pettit, P. (2000). Democracy, electoral and contestatory. In Shapiro, I. and Macedo, S., editors, *Designing Democratic Institutions*, pages 105–144. New York, USA: New York University Press.

[Pirlot et al., 2016] Pirlot, M., Prade, H., and Richard, G. (2016). Completing preferences by means of analogical proportions. In *Modeling Decisions for Artificial Intelligence - 13th International Conference, MDAI 2016, Sant Julià de Lòria, Andorra, September 19-21, 2016. Proceedings*, pages 135–147.

[Prade and Richard, 2018] Prade, H. and Richard, G. (2018). Analogical proportions: From equality to inequality. *Int. J. Approx. Reasoning*, 101:234–254.

[Procaccia, 2018] Procaccia, A. D. (2018). Axioms should explain solutions.

[Rawls, 1971] Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[Ricci et al., 2010] Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B. (2010). *Recommender Systems Handbook*. Springer-Verlag, Berlin, Heidelberg, 1st edition.

[Rolland, 2013] Rolland, A. (2013). Reference-based preferences aggregation procedures in multi-criteria decision making. *European Journal of Operational Research*, 225(3):479 – 486.

[Rossi et al., 2011] Rossi, F., Venable, K. B., and Walsh, T. (2011). *A Short Introduction to Preferences: Between AI and Social Choice*. Morgan & Claypool Publishers, 1st edition.

[Roy, 1978] Roy, B. (1978). ELECTRE III : Un algorithme de classements fondé sur une représentation floue des préférences en présence de critères multiples. *Cahiers du CERO*, 20(1):3–24.

[Roy, 1991] Roy, B. (1991). The outranking approach and the foundations of electre methods. *Theory and decision*, 31(1):49–73.

[Roy, 1993] Roy, B. (1993). Decision science or decision-aid science? *European Journal of Operational Research*, 66(2):184 – 203.

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

[Russell, 1912] Russell, B. (1912). *The problem of Philosophy*. Home University Library.

[Saaty, 1990] Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research*, 48(1):9–26.

[Salo and Hamalainen, 2001] Salo, A. A. and Hamalainen, R. P. (2001). Preference ratios in multiattribute evaluation (prime)-elicitation and decision procedures under incomplete information. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):533–545.

[Savage, 1951] Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67.

[Scott, 1964] Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology*, 1(2):233 – 247.

[Simon, 1991] Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science*, 2(1):125–134.

[Siskos et al., 2005] Siskos, Y., Grigoroudis, E., and Matsatsinis, N. F. (2005). Uta methods. In *Multiple criteria decision analysis: State of the art surveys*, pages 297–334. Springer.

[Slowinski et al., 2002] Slowinski, R., Greco, S., and Matarazzo, B. (2002). Axiomatization of utility, outranking and decision-rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. *Control and Cybernetics*, 31(4):1005–1035.

[Sobrie et al., 2018] Sobrie, O., Gillis, N., Mousseau, V., and Pirlot, M. (2018). Uta-poly and uta-splines: Additive value functions with polynomial marginals. *European Journal of Operational Research*, 264(2):405 – 418.

[Sobrie et al., 2013] Sobrie, O., Mousseau, V., and Pirlot, M. (2013). Learning a majority rule model from large sets of assignment examples. In Perny, P., Pirlot, M., and Tsoukiás, A., editors, *Algorithmic Decision Theory*, volume 8176 of *Lecture Notes in Artificial Intelligence*, pages 336–350. Springer.

[Sobrie et al., 2015] Sobrie, O., Mousseau, V., and Pirlot, M. (2015). Learning the parameters of a non compensatory sorting model. In Walsh, T., editor, *Algorithmic Decision Theory*, volume 9346 of *Lecture Notes in Artificial Intelligence*, pages 153–170, Lexington, KY, USA. Springer.

[Sokolovska et al., 2017] Sokolovska, N., Chevaleyre, Y., Clément, K., and Zucker, J. (2017). The fused lasso penalty for learning interpretable medical scoring systems. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 4504–4511.

[Soos, 2016] Soos, M. (2016). The CryptoMiniSat 5 set of solvers at SAT Competition 2016. *SAT Competition*, 2016:28.

[Spliet and Tervonen, 2014] Spliet, R. and Tervonen, T. (2014). Preference inference with general additive value models and holistic pair-wise statements. *European Journal of Operational Research*, 232(3):607–612.

[Strasser and Antonelli, 2018] Strasser, C. and Antonelli, G. A. (2018). Non-monotonic logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition.

[Tervonen and Figueira, 2008] Tervonen, T. and Figueira, J. R. (2008). A survey on stochastic multicriteria acceptability analysis methods. *Journal of Multi-Criteria Decision Analysis*, 15(1–2):1–14.

[Tintarev, 2007] Tintarev, N. (2007). Explanations of recommendations. In *Proc. ACM conference on Recommender systems*, pages 203–206.

[Tsoukiàs, 2008] Tsoukiàs, A. (2008). From decision theory to decision aiding methodology. *European Journal of Operational Research*, 187:138–161.

[Tversky and Kahneman, 1991] Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):1039–1061.

[Verdú and Poor, 1984] Verdú, S. and Poor, H. V. (1984). On minimax robustness: A general approach and applications. *IEEE Trans. Information Theory*, 30:328–340.

[Vetschera, 2017] Vetschera, R. (2017). Deriving rankings from incomplete preference information: A comparison of different approaches. *European Journal of Operational Research*, 258(1):244 – 253.

[Wachter et al., 2017] Wachter, S., Mittelstadt, B., and Floridi, L. (2017). 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.

[Wakker, 1989] Wakker, P. (1989). *Additive Representations of Preferences: A New Foundation of Decision Analysis*. Theory and Decision Library C. Springer Netherlands.

[Wald, 1950] Wald, A. (1950). *Statistical Decision Functions*. Wiley: New York.

[Walton, 1996] Walton, D. (1996). *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum.

[Walton and Reed, 2002] Walton, D. and Reed, C. (2002). Argumentation schemes and defeasible inferences. In *Workshop on CMNA*.

[Wang and Boutilier, 2003] Wang, T. and Boutilier, C. (2003). Incremental utility elicitation with the minimax regret decision criterion. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, pages 309–318.

[Waterman, 1986] Waterman, D. (1986). *A guide to expert systems*. Addison-Wesley Pub. Co., Reading, MA.

[Wilson, 2009] Wilson, N. (2009). Efficient inference for expressive comparative preference languages. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'19)*, pages 961–966, Pasadena, California, USA.

[Wilson and George, 2017] Wilson, N. and George, A.-M. (2017). Efficient inference and computation of optimal alternatives for preference languages based on lexicographic models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1311–1317.

[Wilson et al., 2015] Wilson, N., George, A.-M., and O'Sullivan, B. (2015). Computation and complexity of preference inference based on hierarchical models. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*, pages 3271–3277. AAAI Press.

[Zopounidis and Doumpos, 2002] Zopounidis, C. and Doumpos, M. (2002). Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, 138(2):229 – 246.

Appendices

A

Proofs

A.1 Proofs of Theorems 2.4 and 2.8

Theorem 2.4. $\mathcal{U}_{\mathcal{P}} \cap \mathcal{N}_{\mathcal{P}} = \varnothing$.

Theorem 2.8. (characterization of necessary preference using covectors) *Given some preference information* $\mathcal{P} \subset \mathbb{P}^2 \subset \mathbb{X}^2$, *and a pair of alternatives* $(x, y) \in \mathbb{X}^2 \setminus \mathcal{U}_{\mathcal{P}}$, *the following propositions are equivalent :*

1. *necessary preference*
$$(x, y) \in \mathcal{N}_{\mathcal{P}}$$

2. *linear feasibility problem*
$$\begin{cases} & (x, y)^\star \times \Delta v < 0 \\ \forall \pi \in \mathcal{P}, & \pi^\star \times \Delta v \geq 0 \\ \forall (i, k) \in \mathbb{I}, & \delta^\star_{(i,k)} \times \Delta v \geq 0 \end{cases} \quad \textit{has no solution } \Delta v \in \mathbb{R}^{\mathbb{I}}$$

3. *combination of statements* $\exists \lambda \in [0, +\infty[^{\mathcal{P}},\ \mu \in [0, +\infty[^{\mathbb{I}}$:
$$(x, y)^\star = \sum_{\pi \in \mathcal{P}} \lambda_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} \mu_{(i,k)} \delta^\star_{(i,k)}$$

4. *integral combination of statements*
$\exists n \in \mathbb{N}^\star,\ \ell \in \mathbb{N}^{\mathcal{P}},\ m \in \mathbb{N}^{\mathbb{I}}$:
$$n\ (x, y)^\star = \sum_{\pi \in \mathcal{P}} \ell_\pi\ \pi^\star + \sum_{(i,k) \in \mathbb{I}} m_{(i,k)}\ \delta^\star_{(i,k)}$$

Proof of (1) $\iff$ (2)

We begin by noticing that

- preference statements involve *core alternatives* (see Def. 2.2.1) : $\mathcal{P} \subset \mathbb{P}$
- pairs of core alternatives are never *unbounded* : $\mathbb{P} \cap \mathcal{U}_{\mathcal{P}} = \varnothing$

Thus, Theorem 2.4 does not apply to core alternatives,so that preference statements, as well as queries between core alternatives, can be represented by covectors. We begin by breaking down the definition of covectors by point of view :

Given a point of view $\forall i \in N$ and two values $\forall x_i, y_i \in \mathbb{X}_i$, let $(x_i, y_i) \in \mathbb{R}^{|\mathbb{P}_i|-1} : \forall k \in \mathbb{N} : 1 \leq k \leq |\mathbb{P}_i| - 1$,

$$(x_i, y_i)^\star_k := \begin{cases} +1, & \text{if } [p_i^k, p_i^{k+1}] \subset [y_i, x_i] \\ -1, & \text{if } [p_i^k, p_i^{k+1}] \cap]x_i, y_i[\neq \varnothing \\ 0, & \text{else.} \end{cases} \tag{A.1}$$

So that, $\forall x, y \in \mathbb{X}, \forall (i,k) \in \mathbb{I},\ (x,y)^\star_{(i,k)} = (x_i, y_i)^\star_k$. By Lemma 2.6, these covectors permit to represent differences of value:

$$\forall i \in N, \forall x_i, y_i \in \mathbb{P}_i, \forall V \in \mathbb{V},\ v_i(x_i) - v_i(y_i) = \sum_{k=1}^{|\mathbb{P}_i|-1} (x_i, y_i)^\star_k \Delta v_{(i,k)} \tag{A.2}$$

In the case of pairs of core alternatives, the objective function as well as the constraints of the minimization problem of Lemma 2.3 can be expressed using covectors and matrix multiplication:

We define the function $\Delta V_{\text{inf}} : \mathbb{X}^2 \to \mathbb{R} \cup \{-\infty\}$ by:

$$\forall x, y \in \mathbb{P},\ \Delta V_{\text{inf}}(x,y) := \inf\ (x,y)^\star \times \Delta v \ \text{ s.t. } \Delta v \in \Omega_{\mathcal{P}} \cap \Omega_{\mathcal{D}}; \tag{A.3}$$

$$\text{with} \quad \Omega_{\mathcal{P}} := \{\Delta v \in \mathbb{R}^{\mathbb{I}} \ :\ \forall \pi \in \mathcal{P},\ \pi^\star \times \Delta v \geq 0\}; \tag{A.4}$$

$$\text{and} \quad \Omega_{\mathcal{D}} := \{\Delta v \in \mathbb{R}^{\mathbb{I}} \ :\ \forall (i,k) \in \mathbb{I},\ \delta^\star_{(i,k)} \times \Delta v \geq 0\}. \tag{A.5}$$

Thus,

$$\forall x, y \in \mathbb{X},\ (x,y) \in \mathcal{N}_{\mathcal{P}} \iff \Delta V_{\text{inf}}(x,y) \geq 0 \tag{A.6}$$

Generally, with alternatives (x, y) not necessarily belonging to the core $\mathbb{P}$, it has been shown [**?**] that minimizing $V(x) - V(y)$ over $V \in \mathbb{V}_{\mathcal{P}}$ is still a linear program, with additional decision variables accounting for the distinct values $\{x_i, y_i\} \notin \mathbb{P}_i$. The $v_i(x_i), v_i(y_i)$ are only constrained by the monotonicity of the marginal value functions, so the problem is separate :

$$\Delta V_{\text{inf}} = \inf_{\Delta v \in \Omega_{\mathcal{P}} \cap \Omega_{\mathcal{D}}} \sum_{i \in N} \inf_{\substack{v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i}} v_i(x_i) - v_i(y_i) \tag{A.7}$$

with, $\forall i \in N$,

$$\begin{cases} UX_i := \{v_i(x_i) \in \mathbb{R} : \forall z_i \in \mathbb{P}_i \cup \{y_i\},\ z_i \succsim_i x_i \Rightarrow v_i(z_i) \geq v_i(x_i)\} \\ LX_i := \{v_i(x_i) \in \mathbb{R} : \forall z_i \in \mathbb{P}_i \cup \{y_i\},\ z_i \precsim_i x_i \Rightarrow v_i(z_i) \leq v_i(x_i)\} \\ UY_i := \{v_i(y_i) \in \mathbb{R} : \forall z_i \in \mathbb{P}_i \cup \{x_i\},\ z_i \succsim_i y_i \Rightarrow v_i(z_i) \geq v_i(y_i)\} \\ LY_i := \{v_i(y_i) \in \mathbb{R} : \forall z_i \in \mathbb{P}_i \cup \{x_i\},\ z_i \precsim_i y_i \Rightarrow v_i(z_i) \leq v_i(l_i)\} \end{cases}$$

Thus, it is possible to circumvent this augmentation of the decision space by :

- considering a given criterion $i \in N$ and a given vector $\Delta v \in \Omega_{\mathcal{P}} \cap \Omega_{\mathcal{D}}$;
- directly assigning the additional decision variables to their optimal values in the inner linear program

$$\inf_{v_i(x_i), v_i(y_i)} v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases};$$

- checking this optimal case is correctly represented, either by an *unbounded pair* or in covector form.

We begin by focusing on the case where the values of $\mathbb{P}_i \cup \{x_i, y_i\}$ are all different. We sort these values in strictly ascending order, and we detail three cases according to the position of x_i and y_i amongst these $|\mathbb{P}_i| + 2$ values :

- the interval $[x_i,\ y_i]$ overflows the set $\mathbb{P}_i$, so that the pair $(x, y) \in \mathcal{U}_{\mathcal{P}}$ is *unbounded*. This case actually encompasses three subcases
  - x_i *has no predecessor*, when x_i is the least element of $\mathbb{P}_i \cup \{x_i, y_i\}$. There is no constraints in $LX_i = \mathbb{R}$;
  - y_i *has no successor*, when y_i is the highest element of $\mathbb{P}_i \cup \{x_i, y_i\}$. There is no constraints in $UY_i = \mathbb{R}$;
  - *both preceding cases are simultaneously satisfied.*

  In any case,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = -\infty,$$

  thus $V_{\inf}(x, y) = -\infty$ and $(x, y) \notin \mathcal{N}_{\mathcal{P}}$, thus proving Theorem 2.4;
- y_i *is the predecessor of* x_i*, so* x_i *is the successor of* y_i. In this case, the constraints UX_i, LX_i, UY_i, LY_i can all be replaced by the single equality $v_i(x_i) = v_i(y_i)$, which defines a solution both feasible and where the objective function is minimized with respect to the decision variables $v_i(x_i), v_i(y_i)$. Meanwhile, we consider the coefficients $(x, y)^{\star}_{(i,k)}$, $1 \leq k < |\mathbb{P}_i|$: the interval $[y_i, x_i]$ does not contain a single core value $p_i^k \in \mathbb{P}_i$,

hence $(x,y)^{\star}_{(i,k)} \neq +1$; the interval $]x_i, y_i[$ is empty, hence $(x,y)^{\star}_{(i,k)} \neq -1$; finally $(x,y)^{\star}_{(i,k)} = 0$. This proves the identity :

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \left\{ \begin{array}{l} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{array} \right. = \sum_{k=1}^{|\mathbb{P}_i|-1} (x,y)^{\star}_{(i,k)} \Delta u_{(i,k)},$$

as both sides are equal to zero.

- *x_i has a predecessor which is not y_i , and y_i has a successor which is not x_i.*

  First, we rewrite $\inf v_i(x_i) - v_i(y_i)$ s.t. $\left\{ \begin{array}{l} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{array} \right.$ as a difference in marginal value between surrogate alternatives in the core $\mathbb{P}_i$.

  The predecessor $\underline{x_i}$ of x_i is given by $\underline{x_i} := \max\{d \in \mathbb{P}_i, d \precsim_i x_i\}$, so that the constraints UX_i, LX_i can both be replaced by the single equality $v_i(x_i) = v_i(\underline{x_i})$, which defines a solution both feasible and where $v_i(x_i)$ is minimal with respect to the decision variable $v_i(x_i)$.

  The successor $\overline{y_i}$ of y_i is given by $\overline{y_i} := \min\{d \in \mathbb{P}_i, d \succsim_i y_i\}$, so that the constraints UY_i, LY_i can both be replaced by the single equality $v_i(y_i) = v_i(\overline{y_i})$, which defines a solution both feasible and where $v_i(y_i)$ is maximal, so the objective function is minimal, with respect to the decision variable $v_i(y_i)$.

  Thus,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \left\{ \begin{array}{l} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{array} \right. = v_i(\underline{x_i}) - v_i(\overline{y_i})$$

  Second, as both surrogate alternatives $\underline{x_i}, \overline{y_i}$ belong to $\mathbb{P}_i$, Lemma 1 ensures that :

$$v_i(\underline{x_i}) - v_i(\overline{y_i}) = \sum_{k=1}^{|\mathbb{P}_i|-1} (\underline{x_i}, \overline{y_i})^{\star}_k \Delta u_{(i,k)}$$

  Third, we check that the covector coefficients for criterion i of the original pair match those of the surrogate pair, that is :

$$\forall k \in \mathbb{N} : 1 \leq k < |\mathbb{P}_i|, \ (x_i, y_i)^{\star}_k = (\underline{x_i}, \overline{y_i})^{\star}_k$$

  The proof is straightforward :

  – if $x_i \succ_i y_i$, then there is at least one attribute value $d \in \mathbb{P}_i$ between x_i and y_i, so that the predecessor of x_i and the successor of y_i are in the

same order, thus $\underline{x_i} \gtrsim_i \overline{y_i}$. Hence, the coefficient indexed by (i,k) of their respective covectors are in $\{0,+1\}$, with value $+1$ respectively when $y_i \lesssim_i p_i^k \prec_i p_i^{k+1} \lesssim_i x_i$ and when $\overline{y_i} \lesssim_i p_i^k \prec_i p_i^{k+1} \lesssim_i \underline{x_i}$. The definition of the surrogate pair ensures these conditions are equivalent.

– if $x_i \prec_i y_i$, then obviously $\underline{x_i} \lesssim_i \overline{y_i}$. Hence, the coefficient of their respective covectors indexed by (i,k) are in $\{0,-1\}$, with value 0 respectively when $y_i \lesssim_i p_i^k$ or $p_i^{k+1} \lesssim_i x_i$, and when $\overline{y_i} \lesssim_i p_i^k$ or $p_i^{k+1} \lesssim_i \underline{x_i}$. The definition of the surrogate pair ensures these conditions are equivalent.

Thus,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = \sum_{k=1}^{|\mathbb{P}_i|-1} (x_i, y_i)_k^\star \Delta u_{(i,k)}$$

The cases where $|\mathbb{P}_i \cup \{x_i, y_i\}| = |\mathbb{P}_i| + 1$ are correctly handled in the discussion above : if overflow (when either $x_i \prec_i \min \mathbb{P}_i$ or $y_i \succ_i \max \mathbb{P}_i$) does not occur, the case $x_i = y_i$ extends the case where the optimal value of $v_i(x_i) - v_i(y_i)$ is zero ; the case where $y_i \in \mathbb{P}_i$ leads to the introduction of $\overline{y_i} := y_i$, and the case where $x_i \in \mathbb{P}_i$ leads to $\underline{x_i} := x_i$.

Finally, for any pair $(x,y) \in \mathbb{X}^2$, we have proven that, in every case, either the pair is unbounded and not in the relation $\mathcal{N}_\mathcal{P}$, or it can be represented by a covector such that:

$$\Delta V_{\inf}(x,y) = \inf_{\Delta v \in \mathbb{R}^{\mathbb{I}}} (x,y)^\star \times \Delta v \text{ s.t. } \begin{cases} \forall \pi \in \mathcal{P},\ \pi^\star \times \Delta v \geq 0 \\ \forall (i,k) \in \mathbb{I},\ \delta^\star_{(i,k)} \times \Delta v \geq 0 \end{cases}$$

□

Proof of (2) ⟺ (3)

By Farkas' lemma, the problem (2) has no solution if, and only if, the objective linear form $(x,y)^\star$ is a linear combination with non-negative coefficients of the constraints linear forms $\{\pi^\star, \pi \in \mathcal{P}\}$ and $\{\delta^\star_{i,k}, (i,k) \in \mathbb{I}\}$. □

Proof of (3) ⟺ (4)

Obviously, (4) ⇒ (3). Conversely, as the covectors involved in (3) have integral coordinates, the non-negative coefficients $\{\lambda_\pi, \pi \in \mathcal{P}\}$ and $\{\mu_{(i,k)}, (i,k) \in \mathbb{I}\}$, if they exist, can be chosen in the field of rational numbers. Multiplying the relation by the common denominator $n \in \mathbb{N}^\star$ of these coefficients leads to (4). □

A.2 Proof of Theorem 3.9

Theorem 3.9 *For any integer p, if there is a subset $A \subseteq N : |A| = 3$ and $\forall i \in A,\ |\mathbb{X}_i| \geq p$, then there is a relation $\mathcal{R}$ satisfying Pareto, transitivity and* first-order cancellation, *and a pair $(x, y) \in \mathcal{R}^{(3)}$ such that $(x, y) \in \mathcal{E}_2(\mathcal{R})$ and any explanation of (x, y) by preference swaps of order at most 2 has a length greater than $2p$.*

For the sketch of the proof, we consider three points of view $N = \{A, B, C\}$ and we construct, for every p, a preference between $x = (0_A, p_B, p_C)$ and $y = ((2p)_A, 0_B, 0_C)$. Starting from alternative $(0_A, p_B, p_C)$, we begin with a preference swap between attributes A and B: $(0_A, p_B) \xrightarrow{(A,B)} (1_A, (p-1)_B)$. Then we perform a preference swap between attributes A and C: $(1_A, p_C) \xrightarrow{(A,C)} (2_A, (p-1)_C)$.We proceed then again by a preference swap between attributes A and B, and so on (the sequence is depicted in Figure A.1).

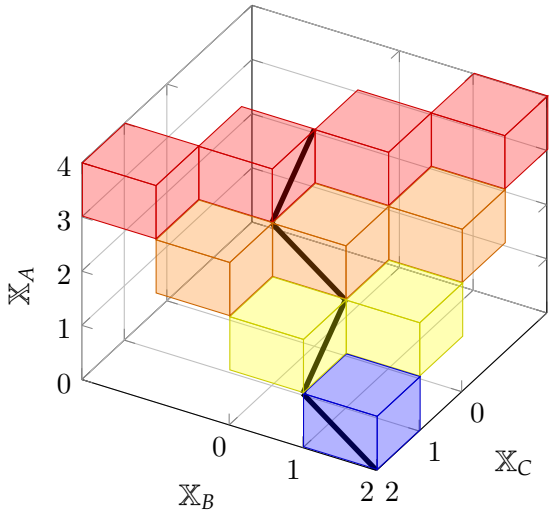

Figure A.1: Description of the sequence

Proof. The proof is based on an instantiation of $\mathcal{R}$ with the necessary preference relation, assuming additive value, inferred from information $\mathcal{P}$, and is denoted by $\mathcal{N}_\mathcal{P}$. Let $p \in \mathbb{N}^*$. Assume that $\mathbb{P}_A := [0..2p] \subseteq \mathbb{X}_A$, $\mathbb{P}_B := [0..p] \subseteq \mathbb{X}_B$ and $\mathbb{P}_C := [0..p] \subseteq \mathbb{X}_C$.Consider the following preference information $\mathcal{P}$, defined by

statements in $(\mathbb{P}_A \times \mathbb{P}_B \times \mathbb{P}_C)^2$ extended ceteris paribus to $\mathbb{X}^2$:

$$\mathcal{P} := \bigcup_{j=0}^{p-1} \{\pi_j, \pi'_j\}, \text{ with} \tag{A.8}$$

$$\pi_j := ((2j)_A, (p-j)_B) \xrightarrow{(A,B)} ((2j+1)_A, (p-j-1)_B); \tag{A.9}$$

$$\pi'_j := ((2j+1)_A, (p-j)_C) \xrightarrow{(A,C)} ((2j+2)_A, (p-j-1)_C). \tag{A.10}$$

We set $x := (0_A, p_B, p_C)$ and $y := ((2p)_A, 0_B, 0_C)$. With this $\mathcal{P}$, we clearly obtain the transitive sequence

$$\Big(x\ ,\ (1_A, (p-1)_B, p_C)\Big)_{\{A,B,C\}} \in \mathcal{P} \qquad \text{(by } \pi_0\text{)};$$

$$\Big((1_A, (p-1)_B, p_C)\ ,\ (2_A, (p-1)_B, (p-1)_C)\Big)_{\{A,B,C\}} \in \mathcal{P} \qquad \text{(by } \pi'_0\text{)};$$

$$\dots$$

$$\Big(((2p-2)_A, 1_B, 1_C)\ ,\ ((2p-1)_A, 0_B, 1_C)\Big)_{\{A,B,C\}} \in \mathcal{P} \qquad \text{(by } \pi_{p-1}\text{)};$$

$$\Big(((2p-1)_A, 0_B, 1_C)\ ,\ ((2p)_A, 0_B, 0_C)\Big)_{\{A,B,C\}} \in \mathcal{P} \qquad \text{(by } \pi'_{p-1}\text{)}.$$

so that $(x, y) \in \mathcal{R}$. This sequence is of length $2\,p$.

There remains to prove that this is the shortest explanation.

We need to to determine $\mathcal{R}^{(}2) \cap (\mathbb{P}_A \times \mathbb{P}_B \times \mathbb{P}_C)^2_{\{A,B,C\}}$.(the other ones can be deduced by Pareto dominance). The preference information $\mathcal{P}$ is very specific. In particular, any value $k \in \mathbb{P}_A \setminus \{0, 2p\}$ appears only in two statements of $\mathcal{P}$:

- if $k = 2j$ is even, on the LHS of π_j and the RHS of π'_{j-1};
- if $k = 2j+1$ is odd, on the RHS of π_j and the LHS of π'_j.

Moreover, we notice that, in the statements π_j and π'_j, the value measuring the fitness according to the point of view A is always increasing from the left hand side to the right hand side, and the values measuring the fitness according to the point of view B and C are decreasing from the left hand side to the right hand side. Hence the elements of Δ_2 cannot be obtained by a combination of two or more preference information. They are obtained only from one preference statement (π_j or pi'_j) and Pareto dominance $\mathcal{D}$. More precisely, $\mathcal{R}^{(2)}$ is composed of the following pairs

$$\Big((i_A, j_B, k_C)\ ,\ (i'_A, j'_B, k'_C)\Big)_{\{A,B,C\}}$$

where either there exists l such that $i_A = 2l$, $i'_A = 2l+1$, $j_B \geq p-l > p-l-1 \geq j'_B$ and $k_C = k'_C$, or there exists l such that $i_A = 2l+1$, $i'_A = 2l+2$, $j_B = j'_B$ and $k_C \geq p-l > p-l-1 \geq k'_C$. From this, one can readily see that the explanation of the preference of x over y described earlier is the shortest one. □

A.3 Proof of Theorem 3.10

Theorem 3.10 (Term by term explanation).
If $\mathcal{P} \subset \mathbb{B}^2, \forall \sigma \in \mathcal{N}_\mathcal{P}$, the following propositions are equivalent :

1. Explainability with a sequence of preference swaps of order at most two: $\sigma \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$
2. Integral combination of statements:
   $\exists a \in \mathbb{N}^\star, \gamma_1, \cdots, \gamma_q \in \widetilde{N}^{(2)}_\mathcal{P}, \ell_1, \cdots, \ell_q \in \mathbb{N}, m_1, \cdots, m_n \in \mathbb{N}$:
   $$a\sigma^\star = \sum_k \ell_k \gamma_k^\star + \sum_k m_k \delta^\star_{(k,1)}$$
3. Reduction to Maximum bipartite matching:
   There is a matching of cardinality $|\sigma^-|$ in the graph of $\widetilde{N}^{(2)}_\mathcal{P} \cap (\sigma^+ \times \sigma^-)$.
4. Term-by-term explanation:
   There is an injection $\phi : \sigma^- \to \sigma^+$ such that $\forall k \in \sigma^-, (\phi(k), k) \in \widetilde{N}^{(2)}_\mathcal{P}$.

We prove Th. 3.10 in four steps : (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (1).

Proof of (1) $\Rightarrow$ (2)

Assume a statement $\sigma := (x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$. By theorem 3.8 and definition 3.4, there is an integer n and a tuple $(e^{(0)}, e^{(1)}, \ldots, e^{(n)}) \in \mathbb{X}^n$ such that $e^{(0)} = x, e^{(n)} = y$ and $(e^{(j)}, e^{(j+1)}) \in \mathcal{D} \cup N^{(2)}_\mathcal{P}$ for any integer $j < n$. This transitive chain of dominance relations and swaps of order 2 can be transformed into the covector relation sought, by induction on the length of the explanation, as described by the following lemmas :

Lemma A.1 (covector representation of dominance relations)**.**

$$\forall \rho \in \mathcal{D}, \exists q \in \{0, +1\}^I : \rho^\star = \sum_{(i,k)\in\mathbb{I}} q_{(i,k)} \delta^\star_{(i,k)}$$

Proof. A dominance relation has no negative argument, so its covector coefficient, given by Equation (2.6), are in $\{0, +1\}$. □

Lemma A.2 (covector representation of transitivity relations)**.**

$$\forall x, y, z \in \mathbb{X}, \exists q \in \mathbb{N}^\mathbb{I} : (x, z)^\star = (x, y)^\star + (y, z)^\star + \sum_{(i,k)\in\mathbb{I}} q_{(i,k)} \delta^\star_{(i,k)}$$

Proof. For core alternatives $x, y, z \in \mathbb{P}$, for any separate value function $V \in \mathbb{V}$,

$$\begin{aligned}(x,z)^\star \times \Delta v =& V(x) - V(z)\\ =& (V(x) - V(y)) + (V(y) - V(z))\\ =& (x,y)^\star \times \Delta v \ + \ (y,z)^\star \times \Delta v\\ =& ((x,y)^\star + (y,z)^\star) \times \Delta v\end{aligned}$$

As the relation above stands for any vector $\Delta v \in [0, +\infty[$, it yields $(x,z)^\star = (x,y)^\star + (y,z)^\star = (x,y)^\star + (y,z)^\star + \sum_{(i,k)\in\mathbb{I}} q_{(i,k)}\delta^\star_{(i,k)}$ with $q = 0$.

For alternatives not necessarily in the core, and for any criterion $i \in N$, the trivial cases where $y_i \in \{x_i, z_i\}$, the case where $x_i = z_i$, or the case where x_i, y_i, z_i are all distinct, divided in 6 subcases considering the order of attributes x_i, y_i, z_i, all lead to $(x,z)^\star_{(i,k)} \geq (x,y)^\star_{(i,k)} + (y,z)^\star_{(i,k)}$ because of the rounding down of broken intervals occuring once in the LHS and twice in the RHS. As both sides are covectors with integer coefficients, the difference $(x,z)^\star - ((x,y)^\star + (y,z)^\star)$ is a covector with non-negative integer coefficients $q_{(i,k)}$. □

Proof of (2) ⇒ (3)

Suppose there exists integer coefficients $a, \ell_1, \cdots, \ell_q$, $m_1, \cdots, m_n$ and preference swaps of order 2 : $\gamma_1, \cdots, \gamma_q$ such that

$$a\sigma^\star = \sum_k \ell_k \gamma_k^\star + \sum_k m_k \delta^\star_{(k,1)} \tag{A.11}$$

Multiplying both sides of the covector equation (A.11) by the vector $(1, \cdots, 1)$, we obtain the relation :

$$M := a(|\sigma^+| - |\sigma^-|) = \sum m_k \geq 0$$

To homogenize the right-hand side, we represent the dominance relation thanks to a dummy criterion : $N' = N \cup \{0\}$ so that a a dominance statement represented by the covector $\delta^\star_{k,1}$ can be represented in the graph of the swap relation between points of view by an edge $(i, 0)$. Defining $\widetilde{\mathcal{D}} := \{(i,0), i \in N\} \subset N'^2$, the relation $\widetilde{\mathcal{D}} \cup \widetilde{N}^{(2)}_{\mathcal{P}}$ is a graph with nodes in N'. Re-indexing coefficients ℓ_k by the positive and negative arguments of swap γ_k (summing up duplicates if needed), and introducing $\ell_{k,0} := m_k$:

$$a\ \sigma^\star = \sum_{\gamma \in \widetilde{\mathcal{D}} \cup \widetilde{N}^{(2)}_{\mathcal{P}}} \ell_{\gamma^+,\gamma^-}\gamma^\star \tag{A.12}$$

In order to complete the flow ℓ, we introduce :

- a source s supplying flow $\ell_{s,i} = a$ to the positive arguments $i \in \sigma^+$;
- a sink t collecting flow $\ell_{j,t} = a$ from the negative arguments $j \in \sigma^-$, and $\ell_{0,t} = M$ from node 0.

Covector equation (A.12) ensures ℓ defines a feasible flow on the graph $(N' \cup \{s,t\}, \widetilde{\mathcal{D}} \cup \widetilde{N}^{(2)}_{\mathcal{P}} \cup (\{s\} \times \sigma^+) \cup (\sigma^- \times \{t\}) \cup \{(0,t)\})$, without capacity constraints, as projection on the i^{th} coordinate ensures flow conservation for node $i \in N'$. Flow ℓ can be decomposed as a superposition of :

- cycles, involving necessary equivalence between the nodes, and not contributing to the value of the flow;
- paths from the source s to the sink t passing through node 0, denoting a dominance relation. Their total contribution to the value of the flow is M;
- paths from the source s to the sink t not passing through node 0, with an overall contribution of $a \times |\sigma^-|$ to the value of the flow. Each of these paths links a positive argument $i_1 \in \sigma^+$ to a negative argument $i_r \in \sigma^-$ through necessary preference swaps of order 2. Transitivity of the necessary preference relation entails that i_1 is necessarily preferred to i_r : the edge (i_1, i_r) belongs to $\widetilde{N}^{(2)}_{\mathcal{P}} \cap (\sigma^+ \times \sigma^-)$.

We reduce the flow ℓ by ignoring the cycles and paths passing through node 0. Also, the flow a carried by the path from source to sink $s \to i_1 \to i_2 \to \ldots \to i_r \to t$ is redirected to edge (i_1, i_r). As a result, we obtain a flow of value $a|\sigma^-|$ on the graph of the relation $\widetilde{N}^{(2)}_{\mathcal{P}}$ restricted to $\sigma^+ \times \sigma^-$. This entails the existence of a matching of cardinality $|\sigma^-|$ in this graph, obtained by setting an upper capacity constraint of value 1 on each edge leaving the source s and entering the sink t (as a cut of capacity C on the network with capacity constraints $c_{i,j} \in \{1, \infty\}$ is a cut of capacity $a \times C$ on the same network with capacity constraints $a \times c_{i,j}$).

Proof of (3) ⇒ (4)

This is simply a rewording.

Proof of (4) ⇒ (1)

Let $\phi : \sigma^- \to \sigma^+$, injective, such that $\forall k \in \sigma^-, (\phi(k), k) \in \widetilde{N}^{(2)}_{\mathcal{P}}$. Given any ordering O of the negative argument set σ^-, we can build a sequence of alternatives of decreasing preference $e^{(0)} := x, e^{(1)}, \ldots, e^{(|\sigma^-|)} \in \mathbb{X}$ such that the k^{th} statement $(e^{(k-1)}, e^{(k)})$ matches the swap between points of view $(\phi(O_k), O_k) \in \widetilde{N}^{(2)}_{\mathcal{P}}$:

$$N^{\neq}_{(e^{(k-1)}, e^{(k)})} := \{\phi(O_k), O_k\} \; ; \; N^{=}_{(e^{(k)}, y)} := N^{=}_{(e^{(k-1)}, y)} \cup \{\phi(O_k), O_k)\}$$

Thus, the sequence of sets $(e^{(k)}, y)^-$ decreases from σ^- to $\varnothing$, one element at a time, and the sequence of sets $(e^{(k)} \succsim y)^+$ also decreases from σ^+ to $\sigma^+ \setminus \phi(\sigma^-)$, one element at a time. If the set $\sigma^+ \setminus \phi(\sigma^-)$ is empty, then $e^{(|\sigma^-|)} = y$, and the sequence $(x = e^{(0)}, \ldots, e^{(|\sigma^-|)} = y)$ is an explanation of $(x, y) \in \mathcal{N}_{\mathcal{P}}$ by preference swaps of order 2, of length $|\sigma^-|$. Else, $e^{(|\sigma^-|)} \neq y$ but $(e^{(|\sigma^-|)}, y)$ is a dominance statement, as its negative argument set is empty. Thus, the sequence $(x = e^{(0)}, e^{(1)}, \ldots, e^{(|\sigma^-|)}, y)$ is an explanation of $(x, y) \in \mathcal{N}_{\mathcal{P}}$ by preference swaps of order 2 and a dominance relation, of length $|\sigma^-| + 1$.

□

Supplementary material

B

Accountable sorting without frontiers

Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)

A Model for Accountable Ordinal Sorting

Khaled Belahcene[1], **Christophe Labreuche**[2], **Nicolas Maudet**[3], **Vincent Mousseau**[1], **Wassila Ouerdane** [1]
[1] LGI, CentraleSupélec, Université Paris-Saclay, Châtenay Malabry, France.
[2]Thales Research & Technology, 91767 Palaiseau Cedex, France.
[3]Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 75005 Paris.
khaled.belahcene, vincent.mousseau, wassila.ouerdane@centralesupelec.fr
christophe.labreuche@thalesgroup.com nicolas.maudet@lip6.fr

Abstract

We address the problem of multicriteria ordinal sorting through the lens of *accountability*, *i.e.* the ability of a human decision-maker to own a recommendation made by the system. We put forward a number of model features that would favor the capability to support the recommendation with a convincing explanation. To account for that, we design a recommender system implementing and formalizing such features. This system outputs explanations under the form of specific *argument schemes* tailored to represent the specific rules of the model. At the end, we discuss possible and promising argumentative perspectives.

1 Introduction

While algorithmic automated decisions or recommendations are nowadays pervasive, there is a growing demand of institutions and citizens to make these recommendations *transparent* and *trustworthy*, while system designers seek *persuasive* recommendations [Tintarev, 2007]. The recent regulation adopted by the European Parliament (known as the General Data Protection Regulation, GDPR) goes further by adding a "right to explanation". According to [Goodman and Flaxman, 2016] *"the GDPR's requirements could require a complete overhaul of standard and widely used algorithmic techniques"*. We interpret this requirement in the strong sense of *accountability*, its litmus test being the ability of the recipient of the recommendation to defend it before other, skeptical, stakeholders of the decision (whereas *trust* requires the recommendation to be consistently accurate, but eventually asks for delegation of the decision to the system; *transparency* simply provides access to the underlying algorithm without concern for technical literacy [Burell, 2016]; and *persuasiveness* is hardly transferable: someone persuaded by a recommendation may not be a good persuader).

Our aim in this paper is thus to build an accountable, ordinal, multicriteria classifier, mapping a *candidate* object to a *recommendation* consisting in one or more categories among a predefined, ordered collection of these. In a multicriteria decision aiding (MCDA) context, the only indisputable relation between objects is the Pareto dominance, occurring when an object outperforms another on all criteria. As the situation is seldom so clear, the rules permitting the comparison of objects need to be enriched, taking into account the knowledge and values of the decision-maker, collected under the label *preference information*, which is also considered as an input of the classifier. We also consider an additional output, an *explanation* aimed at the decision-maker, supporting the recommendation and enabling the accountability sought for. In order to reach this goal of accountability, we make two important assumptions about the recommender system. These *design principles* are as follows:

No jargon. A first step in a MCDA process is to collect decision-maker's preferences information. In order to accurately represent the specific decision process, we opt for an indirect elicitation [Dias *et al.*, 2002]: the decision-maker is never asked any questions about artifacts of the model (e.g. weights). Instead she should express preferences directly in the language of the actual decision situation, i.e. providing direct assignments of typical examples, *reference objects*, to categories.

No arbitrariness. MCDA usually proceeds by representing the reasoning of the decision-maker with a formal parametric model, describing a specific stance. The values of the *preference parameters* are often fitted during an elicitation process, up to a certain point. While many methods proceed by picking a specific, so-called *representative* value of the parameters, we opt for a *robust* approach (to the lack of preference information) [Vincke, 1999; Greco *et al.*, 2008], formulating a –possibly partial – recommendation that cannot be refuted by any judgment function consistent with the preference information.

On top of these principles, we make three further assumptions about the MCDA model, proceeding from the willingness to keep the model accessible to human reasoning.

No compensation. This assumption deals with the interpretation of collected data –the evaluation of objects on various criteria. We assume they are always used comparatively, in a purely ordinal manner: on a given criterion, an alternative is either as good as another one, or strictly worse. Hence, only the *set* of criteria for which an alternative is better is important, regardless of the specific values, and being very good on some criterion cannot compensate for low performance on others. This feature enables the algorithm to proceed without performing any algebraic computation, which makes it particularly suited for explanation. It is shared with established

non-compensatory ordinal sorting models used in the field of MCDA (eg. NCS) [Bouyssou and Marchant, 2007]. Moreover, the use of a 2-valued comparison ($\geq, <$) is similar to [Bouyssou, 1986] rather than [Fishburn, 1976] who proposes a 3-valued one ($<, =, >$).

No values. At the heart of the recommender system is a *preference structure* encoding the comparison of alternatives. There are two main families of structures: those based on *value* [Keeney and Raiffa, 1976], and those based on *outranking relations* [Roy, 1991]. We opt for the latter, as they eschew the construction of a scoring function. An outranking relation naturally provides four outcomes when comparing two alternatives: preference for the former, for the latter, indifference, or incomparability; also, it does not enforce transitivity of preference.

No frontiers. In MCDA, most classifiers link the preference structure and the recommendation of a class by introducing an explicit frontier between classes, defining the limit of each class (a single value for value-based models, a limiting profile for outranking-based ones, e.g. [Leroy *et al.*, 2011]). We do without this construct, as for instance models based on Logical Analysis of Data (LAD) techniques [Crama *et al.*, 1988] which output classification rules. We shall use simple rules permitting to classify a new object by comparing it to a set of already classified *reference objects* (see Sect.2.3).

The general philosophy of these principles must be clear to the reader: accountability should exclude in principle the use of any model artifact that the decision-maker may not properly handle, but at the same time provide enough understanding of the model so as to allow the decision-maker to defend the recommendation *as if it was her own*. Following this, our approach is to enforce these principles by design, and to investigate how far we can get with the resulting sorting model. This approach differs from the recent work of [Ribeiro *et al.*, 2016] which adopts a model-agnostic approach, and builds explanations adapted to virtually any classifier. They obtain extremely promising results in terms of trust. As expected, the explanation cannot be fully faithful to the model (they are "locally" faithful though). It also differs from [Datta *et al.*, 2016] which seeks to extract how influential are input parameters, but keeping a black-box access to the model. While for the trust requirement these approaches are sufficient, our notion of accountability requires to get to grips with the model.

The rest of this paper is as follows. We propose a model implementing and formalizing the different principles, decomposing it in a learning phase (Section 2) and a recommendation phase (Section 3). We provide formal explanations of the recommendation in most cases, in the form of *argument schemes* tailored to represent the specific rules of the model. Section 4 introduces some insights on the description of the sorting problem through an argumentation system. Section 5 concludes the paper, by putting its findings into perspective.

2 Formal Description

In this section, we define a recommender system following the design principles and assumptions, and describe some of its properties.

2.1 The Recommender System

We consider a multicriteria ordinal sorting problem : a collection of objects are evaluated on a set of criteria N. We note $\mathbb{B} := \{0, 1\}$, so that elements of $\mathbb{B}^N$ are at the same time vectors with binary coordinates, and subsets of N, partially ordered by inclusion. The maximal element of $\mathbb{B}^N$ is the unanimous coalition N, also denoted $(1, \ldots, 1)$. The minimal element of $\mathbb{B}^N$ is the empty coalition $\emptyset$, also denoted $(0, \ldots, 0)$. Each criterion $i \in N$ maps an object to a performance value in a totally ordered set $\mathbb{X}_i$, the higher the better. Consequently, each object is described by a performance vector in the partially ordered set $\mathbb{X} = \prod_{i \in N} \mathbb{X}_i$. The objects are to be assigned to some class chosen among an ordered set $\mathbb{K} = \{k_1 \prec \cdots \prec k_p\}$, so that assignment to a class with a high index is desirable.

Formally, let us describe the recommender system as a function mapping a pair $\langle z, \mathcal{P} \rangle$ to a pair $\langle K, \mathcal{E} \rangle$, where:

- The object $z \in \mathbb{X}$ is a *candidate* for sorting;
- $\mathcal{P}$ denotes *preference information* collected from the decision-maker consisting of typical classification examples, a collection of *reference objects* $\mathbb{X}^\star \subset \mathbb{X}$, and their assigned categories $Class : \mathbb{X}^\star \to \mathbb{K}$. For syntactic reasons, we represent it by a set of object-assignment pairs $\mathcal{P} \subset \mathbb{X} \times \mathbb{K}$.

$$\mathcal{P} := \bigcup_{x^\star \in \mathbb{X}^\star} (x^\star, Class(x^\star))$$

- $K \subset \mathbb{K}$ is the *recommendation*, concerning the classes that could be assigned to the candidate (see Sect. 3);
- $\mathcal{E}$ is an *explanation* yet unspecified, supporting the recommendation K (see for instance [Labreuche *et al.*, 2012; Belahcene *et al.*, 2017]), and addressed by Sect. 3.

Example 1. *Objects are evaluated according to four criteria* a, b, c, d *(higher is better). Six reference objects:* $\mathbb{X}^\star := \{A_1, A_2, B_1, B_2, C_1, C_2\}$*, described by the performance table below, are assigned to three classes:* $\mathbb{K} := \{\star \prec \star\star \prec \star\star\star\}$ *and make up the preference information* $\mathcal{P}$*. We consider two candidates:* X, Y *and try to assign them to some possible classes.*

$Object$	a	b	c	d	$Assignment$
A_1	3	3	2.5	0	$\star\star\star$
A_2	3	2	2.1	1	$\star\star\star$
B_1	2	2	1.3	1	$\star\star$
B_2	3	1	3.7	0	$\star\star$
C_1	2	1	1.6	1	$\star$
C_2	1	1	4.1	0	$\star$
X	2	2	1.1	0	?
Y	2	3	1.8	0	?

2.2 The Reasoning of the Decision-Maker

A non-compensatory outranking relation can be represented by a Boolean composite function:

$$\forall x, y \in \mathbb{X},\ x S_\phi y \iff \phi \circ O_N(x, y) = 1$$

where the *observation function* O_N maps a pair of objects to its *concordance set*, and the consistent judgment of the decision-maker, based on these concordance sets, is represented by the *judgment function* ϕ mapping a concordance set to a truth value.

$$O_N : \begin{array}{ccc} \mathbb{X} \times \mathbb{X} & \to & \mathbb{B}^N \\ (x, y) & \mapsto & \{i \in N : x_i \geq y_i\} \end{array}$$

Antecedents of 1 by ϕ, called *true points* in the language of the LAD [Crama *et al.*, 1988], represent *sufficient coalitions of criteria*, while antecedents of 0 by ϕ are *false points* or *insufficient coalitions of criteria*. ϕ is supposed *non-decreasing*, meaning that a superset of a sufficient coalition of criteria is also sufficient, and a subset of an insufficient coalition is also insufficient. Compatibility of the outranking relation S to the Pareto dominance imposes that a unanimous support of criteria is always sufficient, so $\phi(N) = 1$. Conversely, $\phi(\emptyset) = 0$ must hold, so the relation S is not reduced to generalized indifference. Finally, we define the set of any possible judgment function :

$$\phi \in \widehat{\Phi} := \{\phi : \mathbb{B}^N \to \mathbb{B} \ : \phi \nearrow \text{ and } \phi(N) = 1 \text{ and } \phi(\emptyset) = 0\}$$

2.3 Learning From the Assignment Examples

To assign a new object to a category, we shall use the following classification rules:

- *(R1)* an object cannot outrank any object assigned to a strictly better class;
- *(R2)* an object outranks objects assigned to a strictly worse class;
- *(R3)* objects in the same class can be in any position with respect to outranking.

To account for that, we first denote $\succsim_{\mathcal{P}}$ the complete preorder between reference objects induced by $\mathcal{P}$:

$$\left\{ \begin{array}{lcl} x^\star \succsim_{\mathcal{P}} y^\star & \iff & Class(x^\star) \succsim Class(y^\star) \\ x^\star \succ_{\mathcal{P}} y^\star & \iff & Class(x^\star) \succ Class(y^\star) \\ x^\star \sim_{\mathcal{P}} y^\star & \iff & Class(x^\star) = Class(y^\star) \end{array} \right.$$

We consider the strict enforcement of the model rules for reference objects:

- *(R1)* : $\forall x^\star, y^\star \in \mathbb{X}^\star, x^\star \succ_{\mathcal{P}} y^\star \Rightarrow Not\ (y^\star S_\phi x^\star)$;
- *(R2)* : $\forall x^\star, y^\star \in \mathbb{X}^\star, x^\star \succ_{\mathcal{P}} y^\star \Rightarrow x^\star S_\phi y^\star$.

Hence, the assignment of reference objects expressed by $\mathcal{P}$ places upper (by *(R1)*) and lower (by *(R2)*) bounds upon the outranking relation between reference objects. so that:

$$\succ_{\mathcal{P}} \subseteq S_\phi \cap (\mathbb{X}^\star)^2 \subseteq \succsim_{\mathcal{P}}$$

These constraints transfer to the judgment functions. Each pair $(x^\star, y^\star)$ is mapped by the observation function O_N to a coalition of criteria. The observed coalitions $O_N(\mathbb{X}^\star \times \mathbb{X}^\star)$ serve as a learning set for the judgment function ϕ. They are sorted between three sets, yielding necessary conditions on ϕ:

- insufficient coalitions $O_N(\prec_{\mathcal{P}})$ should be mapped to 0;
- sufficient coalitions $O_N(\succ_{\mathcal{P}})$ should be mapped to 1;
- $O_N(\sim_{\mathcal{P}})$, which images by ϕ are not constrained.

Consequently, we define the set $\Phi(\mathcal{P})$ of judgment functions compatible to the preference information $\mathcal{P}$:

$$\Phi(\mathcal{P}) := \{\phi \in \widehat{\Phi} \ : \ \phi \circ O_N(\succ_{\mathcal{P}}) = 1 \text{ and } \phi \circ O_N(\prec_{\mathcal{P}}) = 0\}$$

Example 2. *(ex. 1 continued) In the following table, we detail all the relevant observed coalitions. Sufficient coalitions appear in the upper right side, boldfaced, while insufficient coalitions are in the lower left side.* ***N*** *stands for unanimity, which is self-explanatory.*

	$\star\star\star$		$\star\star$		$\star$	
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	–	–	**abc**	**abd**	**abc**	**abd**
A_2	–	–	**N**	**abd**	**N**	**abd**
B_1	d	bd	–	–	**abd**	**abd**
B_2	acd	ac	–	–	**abc**	**abd**
C_1	d	d	acd	bd	–	–
C_2	cd	c	c	bcd	–	–

2.4 Consistency of Judgment

The set $\Phi(\mathcal{P})$ is empty if, and only if, Pareto dominance is contradicted ($\exists\ x^\star, y^\star \in \mathbb{X}^\star,\ \forall i \in N,\ x_i^\star \geq y_i^\star$ and $Class(x^\star) < Class(y^\star)$), or some coalition of criteria $M \in \mathbb{B}^N$ observed as being sufficient is weaker (for inclusion) than some coalition $M' \in \mathbb{B}^N$ observed as being insufficient. In such a case, we call the preference information $\mathcal{P}$ *inconsistent*; otherwise, it is consistent and $\Phi(\mathcal{P})$ is a *partially defined Boolean function* [Crama *et al.*, 1988]. Combining the constraints on the judgment functions expressed by $\widehat{\Phi}$ and by $\mathcal{P}$, we can compute the true points of $\Phi(\mathcal{P})$. They are the antecedents of 1 common to every judgment function $\phi \in \Phi(\mathcal{P})$, and represent the coalitions *established as sufficient*, by the virtue of being at least as strong as an observed sufficient coalition.

$$\mathcal{T}_{\mathcal{P}} := \{t \in \mathbb{B}^N : \exists\, t_{obs} \in O_N(\succ_{\mathcal{P}}),\ t_{obs} \subseteq t\}$$

Conversely, the false points are the antecedents of zero common to every $\phi \in \Phi(\mathcal{P})$ and represent the coalitions established as insufficient.

$$\mathcal{F}_{\mathcal{P}} := \{f \in \mathbb{B}^N : \exists\, f_{obs} \in O_N(\prec_{\mathcal{P}}),\ f_{obs} \supseteq f\}$$

Proposition 1 details three manners to express inconsistency:

Proposition 1. *For any $\mathcal{P} \subset \mathbb{X} \times \mathbb{K}$, the three following conditions are equivalent and characterize inconsistency:*

1. *Absence of compatible judgment function:* $\Phi(\mathcal{P}) = \emptyset$
2. *Conflicting constraints:* $\mathcal{T}_{\mathcal{P}} \cap \mathcal{F}_{\mathcal{P}} \neq \emptyset$
3. *Explicit contradiction:* $\exists t \in O_N(\succ_{\mathcal{P}}), \exists f \in O_N(\prec_{\mathcal{P}}) : t \subseteq f$

Example 3. *(ex. 2 continued) Coalitions are sorted according to the observations, and monotonicity:*
$O_N(\succ_{\mathcal{P}}) = \{N, abc, abd\} = \mathcal{T}_{\mathcal{P}}$
$O_N(\prec_{\mathcal{P}}) = \{c, d, ac, bd, cd, acd, bcd\}$
$\mathcal{F}_{\mathcal{P}} = \{\emptyset, a, b, c, d, ac, ad, bc, bd, acd, bcd\}$
There is no dispute, as $\mathcal{T}_{\mathcal{P}} \cap \mathcal{F}_{\mathcal{P}} = \emptyset$, but the coalition ab is left undecided.

3 Recommendations and Explanations

In the previous section, we saw how the decision-maker interprets pairwise comparisons between reference objects belonging to different classes as sufficient or insufficient coalitions of criteria. Here comes a new candidate, $z \in \mathbb{X}$. It gauges every reference object in $\mathbb{X}^\star$, yielding $|\mathcal{P}|$ observations $\overrightarrow{o}(z, \mathcal{P}) := \bigcup_{x^\star \in \mathbb{X}^\star} O_N(z, x^\star)$, and is also evaluated by every reference object, yielding $|\mathcal{P}|$ other observations $\overleftarrow{o}(z, \mathcal{P}) := \bigcup_{x^\star \in \mathbb{X}^\star} O_N(x^\star, z)$. Each of these $2|\mathcal{P}|$ observations is interpreted as a *sufficient, insufficient* or *undecided* coalition of criteria.

Example 4. *(ex. 3 continued) The following table augments the one presented in example 2 with the coalitions resulting from comparisons between the reference objects $A_1, A_2, B_1, B_2, C_1, C_2$ and the candidates X, Y.*

	$\star\star\star$		$\star\star$		$\star$		?	?
	A_1	A_2	B_1	B_2	C_1	C_2	X	Y
A_1	–	–	**abc**	**abd**	**abc**	**abd**	***N***	***N***
A_2	–	–	***N***	**abd**	***N***	**abd**	***N***	acd
B_1	d	bd	–	–	**abd**	**abd**	***N***	ad
B_2	acd	ac	–	–	**abc**	**abd**	acd	acd
C_1	d	d	acd	bd	–	–	acd	ad
C_2	cd	c	c	bcd	–	–	cd	cd

X	d	b	(ab)	bd	(ab)	**abd**
Y	bd	b	**abc**	bd	**abc**	**abd**

Non-bracketed coalitions have already been sorted according to the preference information: boldfaced coalitions are those previously established as sufficient, the others are insufficient. Bracketed coalitions are yet undecided. $\forall z \in \{X, Y\}$, $\overrightarrow{o}(z, \mathcal{P})$ appears in the corresponding line, and $\overleftarrow{o}(z, \mathcal{P})$ in the appropriate column.

In this section, we specify the mapping between these observations and the output of the classifier system, the *recommendation* $K(z, \mathcal{P}) \subset \mathbb{K}$ and an *explanation* $\mathcal{E}(k, \mathcal{P})$ supporting it.

3.1 Possible Assignments

As defined by the works of [Greco *et al.*, 2010] about *necessary* and *possible* preference relations, the definition of *possible assignments* is closely related to the notion of *consistency* of an assignment with respect to the corpus of preference information. Defining, as we did in Section 2, $\Phi(\mathcal{P})$ as the set of preference parameters compatible to $\mathcal{P}$, and assuming it is not empty:

- *necessary* assignments are yielded by *every* possible completion of these preference parameters;
- *possible* assignments are yielded by *some* possible completion of these preference parameters;
- *impossible* assignments are yielded by *no* possible completion of these preference parameters;

These sets of assignments are concisely described referring to the set:

$$\widehat{K}(z, \mathcal{P}) := \{k \in \mathbb{K} : \ \Phi(\mathcal{P} \cup \{(z, k)\}) \neq \emptyset\}$$

A possible assignment is in $\widehat{K}(z, \mathcal{P})$, an impossible one is not. When $\widehat{K}(z, \mathcal{P})$ boils down to a singleton, then it is a necessary assignment for z.

This definition of *possible assignment* is straightforward to implement, simply iterating through the set of possible assignments classes $k \in \mathbb{K}$, updating the preference information $\mathcal{P}' \leftarrow \mathcal{P} \cup \{(z, k)\}$, and checking the consistency of $\mathcal{P}'$. Unfortunately, it is a tricky notion when it comes to explaining. The actual unveiling of a Boolean judgment function compatible to the assignment is not very appealing, as it introduces at the same time elements of *jargon* –describing the judgment of the decision-maker as the partition of coalitions of criteria between sufficient and insufficient– and *arbitrariness*, as some coalitions may very well be undecided and should remain so. Consequently, we adopt the following principle: *"Everything is possible, unless proven otherwise"*.

Doing so shifts the burden of proof towards impossibility, focusing on the exhibition of constraints restricting the set $\widehat{K}(z, \mathcal{P})$. We aim at *explaining* these constraints thanks to *statements* of the form $[premises : conclusions]_{scheme}$. We define several *argument schemes*, as formalized by [Walton, 1996] in order to capture stereotypical patterns of human reasoning. These schemes specify the nature and conditions imposed to both premises and conclusions, yielding to valid arguments. We are looking for *complete* explanations, so we must ensure the validity of the implication $premises \Rightarrow conclusions$, and provide *grounded* sets of statements, such that any premise is either the conclusion of another argument, or directly referencing the assumed available information (pairwise comparisons between the reference objects or the candidate, based on criteria or assignment).

In order to make apparent the cause of impossibility, we consider the potential consequences of assigning a candidate to a class through the *additional (in)sufficient coalitions conditional to the assignment of the candidate z to the class k*:

$$\mathbf{\Delta}\mathcal{T}_{\mathcal{P}}(z, k) := \mathcal{T}_{\mathcal{P} \cup \{(z,k)\}} \backslash \mathcal{T}_{\mathcal{P}}; \ \mathbf{\Delta}\mathcal{F}_{\mathcal{P}}(z, k) := \mathcal{F}_{\mathcal{P} \cup \{(z,k)\}} \backslash \mathcal{F}_{\mathcal{P}}$$

We rewrite the impossibility of assigning the candidate z to the class k using the *conflicting constraints* characterization of inconsistency (see Prop. 1). We consider three potential sources of impossibility, sorted by evidence: $\widehat{K}(z, \mathcal{P}) = \bigcap_{i \in \{1,2,3\}} K_i(z, \mathcal{P})$ where:

- $K_1(z, \mathcal{P}) := \{k \in \mathbb{K} : \mathcal{T}_{\mathcal{P}} \cap \Delta\mathcal{F}_{\mathcal{P}}(z, k) = \emptyset\}$ highlights conflicts between established sufficient coalitions, and the assignment of z;
- $K_2(z, \mathcal{P}) := \{k \in \mathbb{K} : \Delta\mathcal{T}_{\mathcal{P}}(z, k) \cap \mathcal{F}_{\mathcal{P}} = \emptyset\}$ highlights conflicts between established insufficient coalitions, and the assignment of z;
- $K_3(z, \mathcal{P}) := \{k \in \mathbb{K} : \Delta\mathcal{T}_{\mathcal{P}}(z, k) \cap \Delta\mathcal{F}_{\mathcal{P}}(z, k) = \emptyset\}$ takes into account the least obvious situation where some assignment of z may be self-contradictory, without conflicting with any previously acknowledged information.

The next section details the impossibilities captured by the set $K_1(z, \mathcal{P})$, and proposes a supporting explanation $\mathcal{E}_1(z, \mathcal{P})$, while the other cases are briefly presented in section 3.3.

3.2 Assignments Contradicting Previously Established Sufficient Coalitions

In this section, we focus on the set $K_1(z, \mathcal{P}) := \{k \in \mathbb{K} : \mathcal{T}_{\mathcal{P}} \cap \Delta\mathcal{F}_{\mathcal{P}}(z, k) = \emptyset\}$. As seen in the previous section this set provides a range of possible assignments for the candidate z, and partially implements the model described by the manifesto exposed in the introduction. We first describe $K_1(z, \mathcal{P})$ as an intersection of constraints, for which we provide a description based on arguments. We prove $K_1(z, \mathcal{P})$ is an interval of $\mathbb{K}$, and provide a short, yet complete, explanation accounting for this recommendation.

For increased readability, we introduce notations for particular sets of classes. For $k \in \mathbb{K}$, let $\mathbb{K}_{\precsim k}$ (resp. $\mathbb{K}_{\succsim k}$) the interval of classes not greater (resp. not lower) than k.

By construction, the recommended set $K_1(z, \mathcal{P})$ is built in order to reject some impossible assignments. To illustrate and understand its behavior, we make up a situation that specifically triggers this rejection flag. Suppose we know that:

(1) the coalition of criteria $T \in \mathbb{B}^N$ is already known to be sufficient, and

(2) the candidate $z \in \mathbb{X}$ is at least as good as the reference object $\underline{x}^\star \in \mathbb{X}^\star$, assigned to class $\underline{k} \in \mathbb{K}$, for all criteria in T.

Then, z outranks $\underline{x}^\star$ and cannot be assigned to a class strictly worse than $\underline{k}$ by application of (R1). This constraint is captured by the set $K_1(z, \mathcal{P})$, as the assignment of z to any class $k \prec \underline{k}$ would lead to conclude that the coalition of criteria $O_N(z, \underline{x}^\star)$ is insufficient, so that the coalition of criteria T would belong to both sets $\Delta\mathcal{F}_\mathcal{P}(z, k)$ and $\mathcal{T}_\mathcal{P}$. Consequently, $k \notin K_1(z, \mathcal{P})$.
If we replace the assumption (2) by:

(2') the reference object $\overline{x}^\star \in \mathbb{X}^\star$, assigned to class $\overline{k} \in \mathbb{K}$, is at least as good as the candidate $z \in \mathbb{X}$ for all criteria in T.

then $\overline{x}^\star \in \mathbb{X}^\star$ outranks z and z cannot be assigned to a class strictly better than $\overline{k}$, as

$$k \succ \overline{k} \Rightarrow \mathcal{T}_\mathcal{P} \ni T \subseteq O_N(\overline{x}^\star, z) \in \Delta\mathcal{F}_\mathcal{P}(z, k) \Rightarrow k \notin K_1(z, \mathcal{P})$$

Reciprocally, any assignment $k_0 \notin K_1(z, \mathcal{P})$ results in a non-empty intersection $\mathcal{T}_\mathcal{P} \cap \Delta\mathcal{F}_\mathcal{P}(z, k_0)$, which involves at least one sufficient coalition $T \in \mathcal{T}_\mathcal{P}$, as in assumption (1), and one stronger, insufficient coalition resulting either from the observations $\overrightarrow{o}(z, \mathcal{P})$, as in assumption (2), or from $\overleftarrow{o}(z, \mathcal{P})$, as in (2').

A statement of type (1) needs to be backed by evidence, so we introduce two argument schemes:

Definition 1. *For any reference objects $a^\star, b^\star \in \mathbb{X}^\star$ and any coalition of criteria $T \in \mathbb{B}^N$, we say the tuple $[a^\star, b^\star : T]_\mathcal{T}$ instantiates the argument scheme* SUFFICIENT COALITION$(\mathcal{P})$ *if, and only if, $T \supseteq O_N(a^\star, b^\star)$ and $a^\star \succ_\mathcal{P} b^\star$. We also say the tuple $[\emptyset : N]_1$ instantiates the argument scheme* WEAK DOMINANCE.

Proposition 2 (Argumentative structure of the sufficient coalitions)**.**

$$\mathcal{T}_\mathcal{P} = \{N\} \cup \bigcup_{[a^\star, b^\star : T]_\mathcal{T}} \{T\}$$

The sufficient coalitions are exactly the conclusions of the arguments instantiating the SUFFICIENT COALITION$(\mathcal{P})$ *scheme.*

In order to account for the atoms of reasoning (2) and (2') and present them to the recipient of the recommendation, we define the corresponding argument schemes.

Definition 2. *For any coalition of criteria $T \in \mathbb{B}^N$, any reference object $x^\star \in \mathbb{X}^\star$ and any class $c \in \mathbb{K}$, we say that:*

- *the tuple $[T, x^\star : \mathbb{K}_{\succsim c}]_{\mathcal{T}/\overrightarrow{o}}$ instantiates the argument scheme* OUTRANKING$(z, \mathcal{P})$ *if, and only if, $T \in \mathcal{T}_\mathcal{P}$ and $\forall i \in T, z_i \geq x^\star_i$ and $class(x^\star) = c$.*
- *the tuple $[T, x^\star : \mathbb{K}_{\precsim c}]_{\mathcal{T}/\overleftarrow{o}}$ instantiates the argument scheme* OUTRANKED$(z, \mathcal{P})$ *if, and only if, $T \in \mathcal{T}_\mathcal{P}$ and $\forall i \in T, x^\star_i \geq z_i$ and $class(x^\star) = c$*

Proposition 3 (Argumentative structure of the recommendation)**.**

$$K_1(z, p) = \mathbb{K} \cap \bigcap_{[T, \underline{x}^\star : \underline{k}]_{\mathcal{T}/\overrightarrow{o}}} \mathbb{K}_{\succsim \underline{k}} \cap \bigcap_{[T, \overline{x}^\star : \overline{k}]_{\mathcal{T}/\overleftarrow{o}}} \mathbb{K}_{\precsim \overline{k}}$$

Proposition 3 is a concise rewording of the necessary and sufficient conditions for a given class *not* to belong to the set $K_1(z, \mathcal{P})$ detailed previously. As a corollary, it shows that $K_1(z, \mathcal{P})$ is an interval of $\mathbb{K}$. Consequently, $K_1(z, \mathcal{P})$ can be completely described by a pair $(\underline{k_1}, \overline{k_1})$ such that:

- $K_1(z, \mathcal{P}) = \mathbb{K}_{\succsim \underline{k_1}} \cap \mathbb{K}_{\precsim \overline{k_1}}$
- the lower bound $\underline{k_1}$ is *maximal*, as there is no class strictly better than $\overline{k_1}$ which is supported by an argument instantiating the OUTRANKING$(z, \mathcal{P})$ scheme. It is *trivial* if $\underline{k_1} = \min \mathbb{K}$ (either when the set OUTRANKING$(z, \overrightarrow{\mathcal{P}})$ is empty, or when it does not support a stronger outcome), in which case it does not need any explanation. If $\underline{k_1} \succ \min \mathbb{K}$, then it admits at least one *explanation* $\underline{E_1}$ composed of an argument $[T, \underline{x}^\star : \mathbb{K}_{\succsim \underline{k_1}}]_{\mathcal{T}/\overrightarrow{o}} \in$ OUTRANKING backed by an argument $[a^\star, b^\star : T]_\mathcal{T} \in$ SUFFICIENT COALITION;
- the upper bound $\overline{k_1}$ is *minimal*, as there is no class strictly worse than $\overline{k_1}$ which is supported by an argument instantiating the OUTRANKED$(z, \mathcal{P})$ scheme. It is *trivial* if $\overline{k_1} = \max \mathbb{K}$, in which case it does not need any explanation. If $\underline{k_1} \prec \max \mathbb{K}$, then it admits at least one *explanation* $\overline{E_1}$ composed of an argument $[T', \overline{x}^\star : \mathbb{K}_{\precsim \overline{k_1}}]_{\mathcal{T}/\overleftarrow{o}} \in$ OUTRANKED backed by an argument $[a^\star, b^\star : T']_\mathcal{T} \in$ SUFFICIENT COALITION.

Finally, the recommended interval $K_1(z, \mathcal{P})$ is supported by an explanation $\mathcal{E}_1$ in the form of a pair $(\underline{E_1}, \overline{E_1})$, where $\underline{E_1}$ and $\overline{E_1}$ can be either the empty set or a pair of arguments. Taken together, all these 0, 2 or 4 arguments prove that any assignment $k \in \mathbb{K} \setminus K_1(z, \mathcal{P})$ should be rejected as "impossible". Such explanation is not necessarily unique, and we denote by $\widehat{\mathcal{E}_1}(z, \mathcal{P})$ the set of suitable explanations.

Example 5. *(ex. 4 continued)*
Using the table presented in Example 4, the set K_1 can be interpreted as "a candidate cannot be assigned a class laying strictly on the right of, nor a class strictly above, a case containing a boldfaced coalition": Consequently,

- $\begin{cases} K_1(X, \mathcal{P}) = \{\star, \star\star\} \\ \mathcal{E}_1(X, \mathcal{P}) \ni (\emptyset, \{[\emptyset : N]_1, [N, B_1 : \precsim \star\star]_{\mathcal{T}/\overleftarrow{o}}\}) \end{cases}$
  X cannot be ranked higher than $\star\star$, because B_1 is rated $\star\star$ and dominates X.
- $\begin{cases} K_1(Y, \mathcal{P}) = \{\star\star, \star\star\star\} \\ \widehat{\mathcal{E}_1}(Y, \mathcal{P}) \ni (\{[A_1, C_1 : \texttt{abc}]_\mathcal{T}, [\texttt{abc}, B_1 : \succsim \star\star]_{\mathcal{T}/\overrightarrow{o}}\}, \emptyset) \end{cases}$
  Y cannot be ranked lower than $\star\star$, because it outranks B_1. Indeed, Y compares to B_1 the same way as A_1 to C_1: it is at least as good on the sufficient coalition of criteria abc.

3.3 Other Impossible Assignments

The set $K_2(z, \mathcal{P})$ is defined symmetrically from $K_1(z, \mathcal{P})$ w.r.t. sufficient and insufficient coalitions. Assignments *not* in $K_2(z, \mathcal{P})$ result from the collision of a coalition of criteria known to be insufficient, and the observation of a candidate object resulting in an even weaker coalition, so outranking is excluded, and all the classes strictly above or below (depending on the direction of observation) the one of the reference object are therefore forbidden. *Mutatis mutandis*, we can define the argument schemes INSUFFICIENT COALITION($\mathcal{P}$), WEAKLY DOMINATED, NOT OUTRANKING($z, \mathcal{P}$), NOT OUTRANKED($z, \mathcal{P}$) and obtain the same structural results, leading to define similar explanations for the lower and upper bounds of the interval $K_2(z, \mathcal{P})$.

Example 6. *(ex. 4 continued)*

Using the table presented in Ex. 4, the set K_2 interprets the insufficient coalitions of the table, those not boldfaced nor parenthesized. A candidate cannot be assigned a class strictly below, nor strictly on the left, of such cases. For instance, $O_N(B_2, X) = \texttt{acd} \in \mathcal{F}_\mathcal{P}$ (e.g. because $O_N(C_1 \prec_\mathcal{P} B_1) = \texttt{acd}$), so X is not outranked by B_2 and should be at least assigned the same class ($\star\star$), and $O_N(X, B_2) = \texttt{bd} \in \mathcal{F}_\mathcal{P}$ (e.g. because it is weaker than $\texttt{bcd} = O_N(C_2 \prec_\mathcal{P} B_2)$), so X does not outrank B_2 and should not be assigned a strictly better class ($\star\star$). In terms of preference, objects X and B_2 are incomparable, and thus should be assigned the same class. Finally, $K_2(X, \mathcal{P}) = \{\star\star\}$.

The set $K_3(z, \mathcal{P})$ excludes inconsistent judgments on yet undecided coalitions of criteria. There is no guarantee that $K_3(z, \mathcal{P})$ has an interval structure. We omit this case due to space limitations.

4 An argumentative Perspective

Along this paper, we proposed the construction of explanations supporting results of a multi-criteria sorting problem, as combinations of arguments schemes. Each instantiation of one of the six previous main schemes (see Def. 1, 2 and their symmetrical forms) provides one type of argument. These arguments may be conflicting, and two different relations can be distinguished:

Conflicting coalitions: we have evidence indicating that a given coalition is potentially at the same time sufficient and insufficient (i.e. there are two coalitions $t \subseteq f$ such that $[a^\star, b^\star : t]_\mathcal{T}$ and $[c^\star, d^\star : f]_\mathcal{F}$). This situation represents an explicit contradiction corresponding to an inconsistency situation (see Sec. 2.4). Such conflicts are not illustrated through the previous examples, however inconsistencies are classical situations within decision problems, as it concerns a human decision-maker.

Conflicting classification: it may occur that, for some candidate, arguments based on the outranking relation point towards an *empty* interval of possible assignments. This situation corresponds to the fact that the sets $K_1(z, \mathcal{P})$ and $K_2(z, \mathcal{P})$ are disjoint, which may happen when either is empty, or when the lower bound of one exceed the upper bound of the other.

Example 7. *(ex. 4 cont.) Y and A_2 are incomparable, Y and B_2 are incomparable, yet A_2 is preferred to B_2. In particular, A_2 ($\star\star\star$) does not outrank Y and Y does not outrank B_2 ($\star\star$) so $K_2(Y, \mathcal{P}) = \emptyset$.*

The impossibility to provide any recommendation is clearly critical from the point of view of decision aiding. These unfortunate situations cannot be ruled out in the general case, as they may stem from Condorcet paradoxes (failures of transitivity) concerning the necessary outranking relation or the necessary not-outranking relation (see e.g. [Köksalan *et al.*, 2009] for a discussion).

The argumentative treatment for our multi-criteria ordinal sorting problem is thus to construct arguments pro and against each possible assignment (of the reference object and the candidate), and to determine among conflicting arguments the *acceptable* ones. This can be done by taking two different perspectives. One way is to rely on the work of [Dung, 1995] - the next question being to identify which semantics are appropriate in our situation. This is close in spirit to an approach presented in [Amgoud and Serrurier, 2007] for classification in *unordered* classes (however in our context the relation between arguments would be symmetric [Coste-Marquis *et al.*, 2005]). Another perspective is to consider the construction of the argumentation system as a dialogue game and to rely on critical questions [Walton, 1996; Ouerdane *et al.*, 2008] to evaluate the arguments. This perspective has the advantage to keep the decision-maker in the loop, which is often essential in a decision situation [Labreuche *et al.*, 2015]. Both approaches look promising and are made possible thanks to the modeling presented in this paper.

5 Conclusion

We have presented a fully accountable multi-criteria ordinal sorting model, based on several design principles and assumptions. The strength of the model is that it solely relies on a simple set of classification rules, which means that each recommendation can be justified by instantiating and combining these rules–nothing else. Several argument schemes have been proposed for that purpose. Interestingly, some of these schemes have a flavour of analogical reasoning, which was studied in the context of classification [Hug *et al.*, 2016]. Now the simplicity of our model comes at a price: there are different situations where inconsistency might occur, and the model is not equipped yet to handle such situations. Facing this issue we can take two stances. The first one is to relax some of our design assumptions. For instance, we may decide that it is actually acceptable for the model to use a *frontier* between classes (allowing to eschew the Condorcet paradox). This would require original explanation techniques to maintain the desired accountability. Another avenue is to handle the inconsistencies thanks to defeasible and non-monotonic reasoning techniques [Brewka *et al.*, 2008]. Our discussion in Section 4 points to formal argumentation as a natural and promising opportunity for future research.

References

[Amgoud and Serrurier, 2007] Leila Amgoud and Mathieu Serrurier. Arguing and explaining classifications. In *Proceeding of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 160, 2007.

[Belahcene *et al.*, 2017] Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183, 2017.

[Bouyssou and Marchant, 2007] Denis Bouyssou and Thierry Marchant. An axiomatic approach to noncompensatory sorting methods in mcdm, i: The case of two categories. *EJOR*, 178(1):217–245, 2007.

[Bouyssou, 1986] Denis Bouyssou. Some remarks on the notion of compensation in mcdm. *EJOR*, 26(1):150–160, 1986.

[Brewka *et al.*, 2008] Gerhard Brewka, Ilkka Niemelä, and Miroslaw Truszczynski. Nonmonotonic reasoning. In *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 239–284. Elsevier, 2008.

[Burell, 2016] Jenna Burell. How the machine "thinks": understanding opacity in machine learning algorithms. *Big Data and Society*, 1(3), 2016.

[Coste-Marquis *et al.*, 2005] Sylvie Coste-Marquis, Caroline Devred, and Pierre Marquis. Symmetric argumentation frameworks. In *Proceedings of the 8th European Conference Symbolic and Quantitative Approaches to Reasoning with Uncertainty ECSQARU*, pages 317–328. Springer, 2005.

[Crama *et al.*, 1988] Yves Crama, Peter L. Hammer, and Toshihide Ibaraki. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research*, 16(1):299–325, 1988.

[Datta *et al.*, 2016] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *The 37th IEEE Symposium on Security and Privacy (Oakland)*, 2016.

[Dias *et al.*, 2002] Luis Dias, Vincent Mousseau, José Figueira, and Joao Clímaco. An aggregation / disaggregation approach to obtain robust conclusions with electre tri. *EJOR*, 138(2):332–348, 2002.

[Dung, 1995] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[Fishburn, 1976] Peter C. Fishburn. Noncompensatory preferences. *Synthese*, 33(2/4):393–403, 1976.

[Goodman and Flaxman, 2016] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". ArXiv e-prints: 1606.08813, June 2016.

[Greco *et al.*, 2008] Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *EJOR*, 191(2):416–436, 2008.

[Greco *et al.*, 2010] Salvatore Greco, Roman Słowinski, José Figueira, and Vincent Mousseau. Robust ordinal regression. In *Trends in Multiple Criteria Decision Analysis*, pages 241–284. Springer Verlag, 2010.

[Hug *et al.*, 2016] Nicolas Hug, Henri Prade, Gilles Richard, and Mathieu Serrurier. Analogical classifiers: A theoretical perspective. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence*, pages 689–697, 2016.

[Keeney and Raiffa, 1976] Ralph L. Keeney and Howard Raiffa. *Decisions with multiple objectives: Preferences and value tradeoffs*. J. Wiley, New York, 1976.

[Köksalan *et al.*, 2009] Murat Köksalan, Vincent Mousseau, Ozgur Ozpeynirci, and Selin Bilgin Ozpeynirci. A new outranking-based approach for assigning alternatives to ordered classes. *Naval Research Logistics*, 56(1):74–85, 2009.

[Labreuche *et al.*, 2012] Christophe Labreuche, Nicolas Maudet, and Wassila Ouerdane. Justifying Dominating Options when Preferential Information is Incomplete. In *ECAI'12*. IOS Press, 2012.

[Labreuche *et al.*, 2015] Christophe Labreuche, Nicolas Maudet, Wassila Ouerdane, and Simon Parsons. A dialogue game for recommendation with adaptive preference models. In *Proceedings of the 14th International Conference on Autonomous Agent and MultiAgent systems (AAMAS)*, pages 959–967, 2015.

[Leroy *et al.*, 2011] Agnes Leroy, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a multiple criteria sorting method. In *ADT*, pages 219–233. Springer, 2011.

[Ouerdane *et al.*, 2008] Wassila Ouerdane, Nicolas Maudet, and Alexis Tsoukiàs. Argument schemes and critical questions for decision aiding process. In *COMMA*, pages 285–296. IOS Press, 2008.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[Roy, 1991] Bernard Roy. The outranking approach and the foundations of electre methods. *Theory and decision*, 31(1):49–73, 1991.

[Tintarev, 2007] Nina Tintarev. Explanations of recommendations. In *Proc. ACM conference on Recommender systems*, pages 203–206, 2007.

[Vincke, 1999] Philippe Vincke. Robust solutions and methods in decision-aid. *Journal of multicriteria decision analysis*, 8(3):181, 1999.

[Walton, 1996] Douglas Walton. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996.

C

Comparing the SAT formulations for Inv-NCS

A new efficient SAT formulation for learning NCS models: numerical results

Khaled Belahcène[1], **Oumaima Khaled**[1], **Vincent Mousseau**[1], **Wassila Ouerdane**[1] and **Ali Tlili**[1]

Abstract. The NonCompensatory Sorting (NCS) model aims at assigning alternatives evaluated on multiple criteria to one of the predefined ordered categories. Computing the NCS parameters from a learning set of assignment examples is computationally demanding. In order to overcome this problem, two formulations based on Boolean satisfiability (SAT) have recently been proposed. The goal of this work is to compare the efficiency of these two formulations. Thus, we first extend the compact formulation to the general case, handling any number of categories, and then representative computational tests are performed.

***Keywords*—** Multicriteria Decision, NonCompensatory Sorting, Preference Elicitation, SAT

Introduction

Multicriteria sorting problems are decision problems in which alternatives evaluated on several criteria should be assigned to one of the ordered predefined categories. Several multicriteria sorting models have been proposed in the literature (see [8] for an overview). Among these multicriteria sorting models, the NonCompensatory Sorting (NCS) model corresponds to a generalization and formal description of the Electre Tri procedure [9]. One of its specificity is to account for the alternative evaluations in an ordinal perspective avoiding compensation, and it also enables to deal meaningfully with qualitative data.

Learning the parameters of NCS from assignment examples (Inv-NCS) aims at computing the parameters of an NCS model, given the desired outputs of the preference aggregation. Solving such a problem is often computationally difficult. Mixed integer linear formulations [10] and heuristic approaches [12, 13] have been proposed for Inv-NCS. Recently, [2] proposed a SAT formulation of this problem which proves to be more efficient than previous approaches. In this paper, we report a second SAT formulation for Inv-NCS [3] described in the context of two categories. We extend this second formulation to the multiple category case, and perform numerical tests to compare the performance of these two SAT formulations.

The paper is organized as follows. Section 1 presents the NCS model. Inv-NCS, the problem of learning the parameters of NCS from assignment examples is defined in Section 2. In Section 3 and 4, we present the two SAT formulations for Inv-NCS. Section 5 describes the empirical test design, the experimental results and discusses these results. A final section groups conclusions and avenues for further research.

1 NonCompensatory Sorting models

This section is devoted to the presentation of the noncompensatory sorting model, introduced in [5, 6].

[1] LGI, CentraleSupélec, Université Paris-Saclay, Gif-Sur-Yvette, France

1.1 Basic notations

Multiple criteria sorting aims at assigning alternatives to one of the predefined ordered categories $C^1, ..., C^p$. The set of alternatives A are evaluated on n criteria, $\mathcal{N} = \{1, 2, \dots, n\}$; hence, an alternative $a \in A$ is characterized by its evaluation vector $(a_1, \dots, a_n)$, with $a_i \in \mathbb{X}_i$ denoting its evaluation on criterion i. Each criterion is equipped $\succsim_i$ with a weak preference relation 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.

1.2 Sorting with a single profile

In the NonCompensatory Sorting model NCS, the boundaries between categories are defined by profiles. Therefore, a single profile corresponds to the case where alternatives are sorted between two ordered categories that we label as GOOD and BAD. A specific sorting procedure is described by a pair of parameters:

- 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, required to be an upset of the power set of the criteria.

These notions are combined into the following assignment rule:

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

An alternative is considered as GOOD if, and only if, it is better than the limiting profile b according to a sufficient coalition of criteria.

1.3 Sorting into multiple categories

With p categories, the parameter space is extended accordingly, with approved sets $\langle \mathcal{A}_i^k \rangle_{i \in \mathcal{N},\ k \in [2..p]}$ and sufficient coalitions $\langle \mathcal{T}^k \rangle_{k \in [2..p]}$ declined per boundary.

The ordering of the categories $\{C^1 \prec \dots \prec C^p\}$ 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 $\mathcal{T}^2 \supseteq \dots \supseteq \mathcal{T}^p$, and also a nesting of the approved sets: $\forall i \in \mathcal{N}, \forall k \in [2..p]$, $\mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \precsim_i)$ and $\mathcal{A}_i^2 \supseteq \dots \supseteq \mathcal{A}_i^p$.

These tuples of parameters are augmented on both ends with trivial values: $\mathcal{T}^1 = \mathcal{P}(\mathcal{N})$, $\mathcal{T}^{p+1} = \emptyset$, and $\forall i \in \mathcal{N}$, $\mathcal{A}_i^1 = \mathbb{X}$, $\mathcal{A}_i^{p+1} = \emptyset$. With $\omega = (\langle\ \mathcal{A}_i^k\ \rangle_{i \in \mathcal{N},\ k \in [2..p]}, \langle\ \mathcal{T}^k\ \rangle_{k \in [2..p]})$, [6] defines the sorting function NCS_ω from $\mathbb{X}$ to $\{C^1 \prec \dots \prec C^p\}$ with the *noncompensatory sorting rule*:

$$NCS_\omega(x) = C^k \Leftrightarrow \begin{cases} \{i \in \mathcal{N} : x \in \mathcal{A}_i^k\} & \in \mathcal{T}^k \\ \text{and } \{i \in \mathcal{N} : x \in \mathcal{A}_i^{k+1}\} & \notin \mathcal{T}^{k+1} \end{cases} \quad (1)$$

1.4 An illustrative example

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

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 1. Performance table

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

- The attributes of each model are sorted between average ($\star$/ ■), good ($\star\star$/ ■) and excellent ($\star\star\star$/ ■) by comparison to the profiles given in Table 2. The resulting labeling of the six alternatives according to each criterion is depicted in Figure 1 and Table 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. Limiting profiles

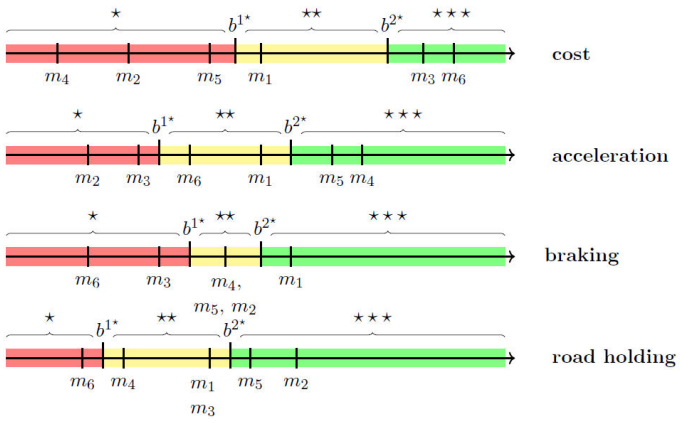

Figure 1. Representation of performances w.r.t. category limits

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 3. 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 noncompensatory model. Sufficient coalitions are represented on Figure 2. Finally, the model yields the assignments presented in Table 4.

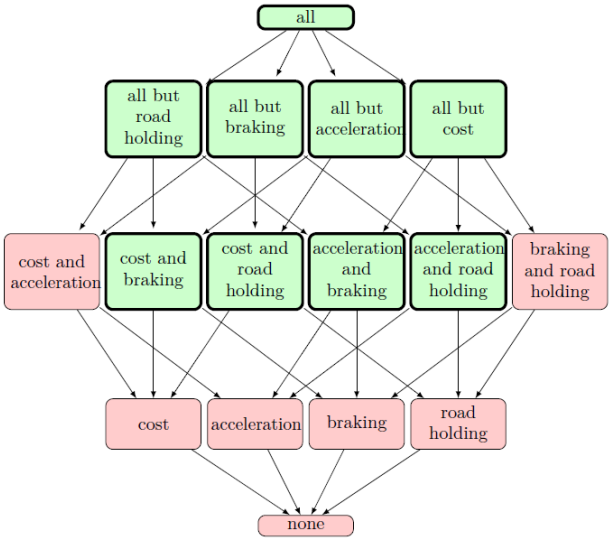

Figure 2. Sufficient (green/thick-bordered) and insufficient (red/thin-bordered) coalitions of criteria. Arrows denote coalition strength

model	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 4. Model Assignments

1.5 Variants of the NCS Model

In this section, we mention a number of variants of the noncompensatory sorting model that can be found in the literature; these variants correspond to simplifications of the model, with additional assumptions that restrict the parameters—limiting profiles and sufficient coalitions—either explicitly or implicitly.

The set of preference parameters – all the pairs $(\langle b \rangle, \langle \mathcal{T} \rangle)$ can be considered too wide and too unwieldy for practical use in the context of a decision aiding process. Therefore, following [6], one may consider to explicitly restrict either the sequence of limiting profiles, or the sequence of sufficient coalitions:

- *noncompensatory sorting with a unique set of sufficient coalitions*: $\mathcal{T}^2 = \cdots = \mathcal{T}^p$;
- *noncompensatory sorting with a unique limiting profile* $b^2 = \cdots = b^p$ or, equivalently, $\forall i \in \mathcal{N},\ \mathcal{A}_i^2 = \cdots = \mathcal{A}_i^p$.

k-additive representations of sufficient coalitions The explicit representation of sufficient coalitions can either use a boolean flag for each subset of $\mathcal{N}$, or, more economically, only mention an antichain of minimally sufficient subsets. Whatever the chosen representation, it is not *compact*, as it potentially demands to store an amount of information that is exponential in the number of criteria.

A simplifying assumption consists in representing sufficient coalitions in an analogy to a voting setting: each criterion $i \in \mathcal{N}$ is assigned a *voting power* $w_i \geq 0$ so that a given coalition of criteria $B \subseteq \mathcal{N}$ is deemed sufficient if, and only if, its combined voting power $\sum_{i \in B} w_i$ is greater than a given *qualification threshold* λ.

$$\exists \lambda, \langle w_i \rangle_{i \in \mathcal{N}} \in [0, +\infty[:\ \forall B \subseteq \mathcal{N}, B \in \mathcal{T} \iff \sum_{i \in \mathcal{N}} w_i \geq \lambda. \tag{2}$$

With this rule, the sufficient coalitions are represented in a compact form which is more amenable to linear programming.

The majority rule is used to represent sufficient coalitions both in Electre-Tri [11] and most variants of the MR-Sort model [10]. As a remarkable exception, [13] considers an extension of the majority rule where the voting power of criteria are replaced by a *capacity*, a function mapping criteria coalitions to nonnegative real numbers that is nondecreasing w.r.t. set inclusion. A coalition is considered sufficient if, and only if, its capacity exceeds the qualification threshold. This variant of the MR-Sort model is obviously equivalent to the NCS model, it offers two additional features:

- a numeric representation for the NCS model, that remains linear and can be handled by techniques and pieces of software dedicated to linear algebra, such as mixed integer programming (MIP) solvers;
- the notion of using a general capacity can easily be restricted to using limited forms, called k-capacities that restrict interactions between criteria. As 1-capacities boil down to the additive form of the majority rule, the normative assumption of representing upsets with k-capacities is often called *k-additivity*. This nesting of assumptions, with $k = 1$ corresponding to the majority rule and $k = |\mathcal{N}|$ corresponding to the general NCS model, goes along a progressive expansion of the numeric parameter space, as going from $(k-1)$ to k additivity requires k among $|\mathcal{N}|$ additional parameters.

2 Learning an NCS model

For a given decision situation, assuming the NCS model is relevant to structure the decision maker's preferences, what should be the parameters values to fully specify the NCS model that corresponds to the decision maker(DM) viewpoint? An option would be to simply ask the decision maker to describe, to her best knowledge, the limit profiles between categories and to enumerate the minimal sufficient coalitions. In order to get this information as quickly and reliably as possible, an analyst could make good use of the *model-based elicitation strategy* described in [4], as it permits to obtain these parameters by asking the decision maker to only provide holistic preference judgment – should some (fictitious) alternative be assigned to some category – and builds the shortest questionnaire.

We opt for a more indirect setup, close to a machine learning paradigm, where a set of reference assignments is given and assumed to describe the decision maker's point of view, and the aim is to *extend* these assignments with a NCS model. In this context, we usually refer to an *assignment* as a function mapping a subset of *reference alternatives* $\mathbb{X}^\star \subset \mathbb{X}$ to the ordered set of categories $C^1 \prec \cdots \prec C^p$. These reference alternatives highlight values of interest on each criterion $i \in \mathcal{N}$, $\mathbb{X}_i^\star = \bigcup_{x \in \mathbb{X}^\star} \{x_i\}$. We are looking for suitable preference parameters specifying a noncompensatory sorting model.

cinque

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

- $\mathcal{N}$ is a set of criteria;
- $\mathbb{X}$ is a set of *alternatives*;
- $\langle \succsim_i \rangle_{i \in N} \in \mathbb{X}^2$ 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* ordered by *exigence*;
- $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$ is an *assignment* of the reference alternatives to the categories.

When referring to an instance, we often shorten this sextuple as 'α'.

Parameters. Given a context, a *parameter* ω of the NCS model is a couple $(\langle\ \mathcal{A}_i^k\ \rangle_{i \in \mathcal{N},\ k \in [2..p]}, \langle\ \mathcal{T}^k\ \rangle_{k \in [2..p]})$, where the *sufficient coalitions* satisfy: $\forall k \in [2..p],\ \mathcal{T}^k$ is an upset of $(2^{\mathcal{N}}, \subseteq)$, and $\mathcal{T}^2 \supseteq \cdots \supseteq \mathcal{T}^p$; and the *approved sets* satisfy $\forall i \in \mathcal{N}$, $\forall k \in [2..p],\ \mathcal{A}_i^k$ is an upset of $(\mathbb{X}_i, \succsim_i)$ and $\mathcal{A}_i^2 \supseteq \cdots \supseteq \mathcal{A}_i^p$.

Sorting rule. Given a parameter $\omega = (\langle\ \mathcal{A}_i^k\ \rangle_{i \in \mathcal{N},\ k \in [2..p]}, \langle\ \mathcal{T}^k\ \rangle_{k \in [2..p]})$, augmented with trivial values $\mathcal{T}^1 := \mathcal{P}(\mathcal{N})$, $T^{p+1} := \emptyset$, $\forall i \in \mathcal{N},\ \mathcal{A}_i^2 = \mathbb{X},\ \mathcal{A}_i^{p+1} = \emptyset$, NCS_ω is the function from $\mathbb{X}$ to $\{C^1 \prec \ldots \prec C^p\}$ satisfying:

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

Solutions. Given a context, a *solution* of the instance α of the Inv-NCS problem is a parameter ω of the NCS model such that $\forall x \in \mathbb{X}^\star,\ \alpha(x) = NCS_\omega(x)$.

Throughout this paper, we assume the expression of preference is free of noise. We are only interested in determining if the given assignment can be represented in the noncompensatory sorting model.

3 A first SAT formulation for Inv-NCS based on coalitions

This section describes and extends a SAT formulation for Inv-NCS initially given in [2]. We provide here an informal presentation of the approach; formal justification can be found in [2]. The formulation Φ_α^C yielded by the encoding presented in this section is based on an explicit representation of the parameter space of the noncompensatory sorting model—the pairs are composed of a sequence of approved sets and a sequence of sufficient coalitions.

Variables. The Boolean function Φ_α^C operates on two types of variables:

- 'a' variables, indexed by a criterion $i \in \mathcal{N}$, an exigence level $k \in [2..p]$ and a reference value $x \in \mathbb{X}^\star$, represent the approved sets $\mathcal{A}_i$, with the following semantic: $a_{i,k,x} = 1 \Leftrightarrow x \in \mathcal{A}_i^k$ *i.e.* x is approved at level k according to i;

- 't' variables, indexed by a coalition of criteria $B \subset \mathcal{N}$ and an exigence level $k \in [2..p]$, represent the sufficient coalitions $\mathcal{T}$, with the following semantic: $t_{B,k} = 1 \Leftrightarrow B \in \mathcal{T}^k$ i.e. the coalition B is sufficient at level k;

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

- Clauses ϕ^{C1}_α ensure that each approved set $\mathcal{A}^k_i$ is an upset of $(\mathbb{X}^\star, \succsim_i)$: if for a criterion i and an exigence value k, the value x is approved, then any value $x' \succsim_i x$ must also be approved.
- Clauses ϕ^{C2}_α ensure that approved sets are ordered by set inclusion according to their exigence level: if an alternative x is approved at exigence level k according to the criterion i, it should also be approved at exigence level $k' < k$.
- Clauses ϕ^{C3}_α ensure that each set of sufficient coalitions $\mathcal{T}$ is an upset for inclusion: if a coalition B is deemed sufficient at exigence level k, then a stronger coalition $B' \supset B$ should also be deemed sufficient at this level.
- Clauses ϕ^{C4}_α ensure that set of sufficient coalitions are ordered by inclusion according to their exigence level: if a coalition B is deemed insufficient at exigence level k, it should also be at any level $k' > k$.
- Clauses ϕ^{C5}_α ensure that each alternative is not approved by a sufficient coalition of criteria at an exigence level above the one corresponding to its assigned category.
- Clauses ϕ^{C6}_α ensure that each alternative is approved by a sufficient coalition of criteria at an exigence level corresponding to its assignment.

Definition 3.1. *Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$, the boolean function Φ^C_α with variables $\langle a_{i,k,x}\rangle_{i\in\mathcal{N},\ k\in[2..p],\ x\in\mathbb{X}^\star}$ and $\langle t_{B,k}\rangle_{B\subseteq\mathcal{N},\ k\in[2..p]}$, as the conjunction of clauses is defined:*

$$\Phi^C_\alpha = \phi^{C1}_\alpha \wedge \phi^{C2}_\alpha \wedge \phi^{C3}_\alpha \wedge \phi^{C4}_\alpha \wedge \phi^{C5}_\alpha \wedge \phi^{C6}_\alpha$$

$$\begin{aligned}
\phi^{C1}_\alpha &= \bigwedge_{i\in\mathcal{N},\ k\in[2..p]} \bigwedge_{x' \succsim_i x\ \in\mathbb{X}^\star} && (a_{i,k,x'} \vee \neg a_{i,k,x})\\
\phi^{C2}_\alpha &= \bigwedge_{i\in\mathcal{N},\ k<k'\in[2..p],\ x\in\mathbb{X}^\star} && (a_{i,k,x} \vee \neg a_{i,k',x})\\
\phi^{C3}_\alpha &= \bigwedge_{B\subset B'\subseteq\mathcal{N},\ k\in[2..p]} && (t_{B',k} \vee \neg t_{B,k})\\
\phi^{C4}_\alpha &= \bigwedge_{B\subseteq\mathcal{N},\ k<k'\in[2..p]} && (t_{B,k} \vee \neg t_{B,k'})\\
\phi^{C5}_\alpha &= \bigwedge_{B\subseteq\mathcal{N},\ k\in[2..p]} \bigwedge_{x\in\alpha^{-1}(C^{k-1})} && (\bigvee_{i\in B} \neg a_{i,k,x} \vee \neg t_{B,k})\\
\phi^{C6}_\alpha &= \bigwedge_{B\subseteq\mathcal{N},\ k\in[2..p]} \bigwedge_{x\in\alpha^{-1}(C^{k})} && (\bigvee_{i\in B} a_{i,k,x} \vee t_{\mathcal{N}\setminus B,k})
\end{aligned}$$

Clauses ϕ^{C1}_α ensure monotonicity of approved sets $\mathcal{A}_k$ with respect to the evaluations on each criterion; ϕ^{C3}_α force the set of sufficient coalitions to be compatible with inclusion; ϕ^{C2}_α and ϕ^{C4}_α enforce that approved sets $\mathcal{A}_k$ and sets of sufficient coalitions get stronger for higher exigence levels; clauses ϕ^{C5}_α and ϕ^{C6}_α ensure that reference alternatives are correctly assigned ($NCS_\omega \equiv \alpha$).

Written as such, they are highly redundant, threatening computational efficiency. Instead, it is sufficient to consider clauses where ordered elements are adjacent to each other.

Model variants. As discussed in Section 1.5, the NCS model has many variants. Φ^C_α can easily be modified to account for two popular restrictions of the model:

- unique profiles—drop the index k concerning the exigence level for the 'a' variables, ignore the conjunction over exigence levels for clauses ϕ^{C1}_α, and ignore clauses ϕ^{C2}_α altogether;
- unique set of sufficient coalitions—drop the index k concerning the exigence level for the 't' variables, ignore the conjunction over exigence levels for clauses ϕ^{C3}_α and ignore clauses, ϕ^{C4}_α altogether.

4 A second SAT formulation based on pairwise separation conditions

The boolean satisfiability formulation presented in this section, denoted Φ^P_α, was originally described in [3] but only focusing on the case with two categories $C^1 \equiv$ BAD $\prec C^2 \equiv$ GOOD . We extend this formulation to the general case, with any number of categories. This second formulation is more compact than the first one as it handles explicitly a set of sufficient coalitions that lies in the power set of the criteria.

Encoding. Similarly to the formulation Φ^C_α described in Section 3, the formulation Φ^P_α operates on two types of variables:

- 'a' variables, representing the approved sets, with the exact same semantics as their counterpart in Φ^C_α, i.e.

$$a_{i,x} = \begin{cases} 1 & \text{if } x \in \mathcal{A}_i \text{ i.e. } x \text{ is approved according to } i;\\ 0 & \text{else.}\end{cases}$$

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

$$s_{i,g,b} = \begin{cases} 1 & \text{if } g \in \mathcal{A}_i \text{ and } b \notin \mathcal{A}_i;\\ 0 & \text{else.}\end{cases}$$

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

Definition 4.1. *Given an instance of Inv-NCS with two categories and an assignment $\alpha : \mathbb{X}^\star \to \{$ BAD $\prec$ GOOD $\}$, we define the boolean function Φ^P_α with variables $\langle a_{i,x}\rangle_{i\in\mathcal{N},\ x\in\mathbb{X}^\star}$ and $\langle s_{i,g,b}\rangle_{i\in\mathcal{N},\ g\in\alpha^{-1}(\text{ GOOD }),\ b\in\alpha^{-1}(\text{ BAD })}$, as the conjunction of clauses:*

$$\Phi^P_\alpha = \phi^{P1}_\alpha \wedge \phi^{P2}_\alpha \wedge \phi^{P3}_\alpha \wedge \phi^{P4}_\alpha$$

$$\begin{aligned}
\phi^{P1}_\alpha &= \bigwedge_{i\in\mathcal{N}} \bigwedge_{x'\succsim_i x\in\mathbb{X}^\star} (a_{i,x'} \vee \neg a_{i,x})\\
\phi^{P2}_\alpha &= \bigwedge_{i\in\mathcal{N},\ g\in\alpha^{-1}(\text{ GOOD }),\ b\in\alpha^{-1}(\text{ BAD })} (\neg s_{i,g,b} \vee \neg a_{i,b})\\
\phi^{P3}_\alpha &= \bigwedge_{i\in\mathcal{N},\ g\in\alpha^{-1}(\text{ GOOD }),\ b\in\alpha^{-1}(\text{ BAD })} (\neg s_{i,g,b} \vee a_{i,g})\\
\phi^{P4}_\alpha &= \bigwedge_{g\in\alpha^{-1}(\text{ GOOD }),\ b\in\alpha^{-1}(\text{ BAD })} (\bigvee_{i\in\mathcal{N}} s_{i,g,b})
\end{aligned}$$

The formulation is compact: $O(|\mathcal{N}|.|\mathbb{X}|^2)$ variables, $O(|\mathcal{N}|.|\mathbb{X}|^2)$ binary clauses and $O(|\mathbb{X}|^2)$ $|\mathcal{N}|$-ary clauses, whereas the number of 't' variables in the first formulation increases exponentially with the number of criteria.

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

More than two categories The case where there are $p > 2$ categories $\{C^1 \prec \ldots \prec C^p\}$ requires a few adaptations of the formulation given in the preceding section, and presented in [3]. It relies mostly on the fact that an NCS model with p categories is informally the combination of $p-1$ NCS models with two categories whose parameters satisfy the nesting conditions on coalitions and satisfactory values. Given an assignment α and an exigence level $k \in [2..p]$, we define the set of alternatives assigned at least to C^k as

$$C^{\succeq k} = \bigcup_{k' \in [k..p]} C^{k'}.$$

We propose the following definition for $\Phi_\alpha^{P'}$, that coincides with Φ_α^P (see Definition 4.1) when $p = 2$.

Definition 4.2. *Given an instance of Inv-NCS with an assignment $\alpha : \mathbb{X}^\star \to \{C^1 \prec \ldots \prec C^p\}$, we define the boolean function $\Phi_\alpha^{P'}$ with variables $\langle a_{i,k,x}\rangle_{i\in\mathcal{N},\ k\in[2..p],x\in\mathbb{X}^\star}$ and $\langle s_{i,k,g,b}\rangle_{i\in\mathcal{N},\ k\in[2..p],\ g\in\alpha^{-1}(G^{\succeq k}),\ b\notin\alpha^{-1}(G^{\succeq k})}$, as the conjunction of clauses:*

$$\Phi_\alpha^{P'} = \phi_\alpha^{P'1} \wedge \phi_\alpha^{P'2} \wedge \phi_\alpha^{P'3} \wedge \phi_\alpha^{P'4} \wedge \phi_\alpha^{P'5}$$

$$\phi_\alpha^{P'1} = \bigwedge_{i\in\mathcal{N},\ k\in[2..p]} \bigwedge_{x'\succsim_i x\in\mathbb{X}^\star} (a_{i,k,x'} \vee \neg a_{i,k,x})$$

$$\phi_\alpha^{P'2} = \bigwedge_{i\in\mathcal{N},\ k<k'\in[2..p],\ x\in\mathbb{X}^\star} (a_{i,k,x} \vee \neg a_{i,k',x})$$

$$\phi_\alpha^{P'3} = \bigwedge_{i\in\mathcal{N},\ k\in[2..p]} \bigwedge_{g\in\alpha^{-1}(G^{\succeq k}),\ b\notin\alpha^{-1}(G^{\succeq k})} (\neg s_{i,k,g,b} \vee \neg a_{i,k,b})$$

$$\phi_\alpha^{P'4} = \bigwedge_{i\in\mathcal{N},\ k\in[2..p]} \bigwedge_{g\in\alpha^{-1}(G^{\succeq k}),b\notin\alpha^{-1}(G^{\succeq k})} (\neg s_{i,k,g,b} \vee a_{i,k,g})$$

$$\phi_\alpha^{P'5} = \bigwedge_{k\in[2..p]\ g\in\alpha^{-1}(G^{\succeq k}),\ b\notin\alpha^{-1}(G^{\succeq k})} (\bigvee_{i\in\mathcal{N}} s_{i,k,g,b})$$

The remarks made about an efficient implementation of Φ_α^C apply here: many clauses are redundant in $\phi_\alpha^{P'1}$ and $\phi_\alpha^{P'2}$ and can safely be ignored. The remark concerning the non-uniqueness of $\mathcal{T}$ in the case of two categories also applies for more than two categories to $\mathcal{T}^k$ which are not uniquely defined by $\Phi_\alpha^{P'}$.

5 Computational study

In this section, we present an empirical study that evaluates the comparative performance of the two SAT formulations presented in Sections 3 and 4 in order to assess their relative advantages. Note that performances of SAT^C (Section 3) have already been proved to be superior to MIP approaches by [2].

We restrict our experimental study to the case of U-NCS (noncompensatory sorting with a unique set of sufficient coalitions) model, i.e., where $\mathcal{T}^2 = \cdots = \mathcal{T}^p$. For both SAT formulations, we solve instances of the problem of learning a U-NCS model given the assignment of a set of reference alternatives using a state-of-the-art SAT solver.

5.1 Experimental design

Our experiments take as input the assignment of a set of alternatives $\mathbb{X}^*$ (described by tuples of evaluations on a set of criteria $\mathcal{N}$), to a set of categories $C^1 \prec \ldots \prec C^p$. The computing performance is measured in practice, by solving actual instances of the problem and reporting the computation time required by the solver. This experimental study is run on a laptop with Windows 10 (64 bit) equipped with an Intel(R) Xeon(R) CPU E5-1620 v4 @3.5GHz and 16 GB of RAM.

Dataset generation:

In this paper, we only consider carefully crafted random dataset as an input. On the one hand, the implementation of both formulations is not yet equipped with the capability to deal with noisy inputs, so we do not consider feeding them with actual preference data. On the other hand, using totally random and unstructured inputs makes little sense in the context of algorithmic decision. In order to ensure the preference data which makes sense, we use a decision model to generate it, and, in particular, a model compatible with the noncompensatory instance we are postulating. Precisely, we use an MR-Sort model for generating the learning set, a model that particularizes NCS and U-NCS by postulating the set of sufficient coalitions which possess an additive structure (see Section 1). This choice ensures that both formulations should succeed in finding the parameters of a model extending the reference assignment.

When generating a dataset, we consider the number of criteria $|\mathcal{N}|$, the number of categories p, and the number of reference alternatives $|\mathbb{X}^*|$ as parameters. All datasets have been tested considering a baseline configuration composed of: 3 categories, 9 criteria, and 128 reference alternatives. Other configurations (triplet of parameters) are also tested while deviating from the baseline on a single parameter either the number of categories p from 2 to 5, the number of criteria, among $\{3, 5, 7, 9, 11\}$, or the number of reference alternative among $\{16, 32, 64, 128, 256, 512, 1024\}$.

For each triplet of parameters, we sample 100 MR-Sort M^0, and record the computation time needed to obtain an NCS Model for SAT^C and SAT^P and the restoration rate for both formulations.

We consider all criteria take continuous value in the interval [0,1], which is computationally more demanding than the case where one criterion has a finite set of values. A set of ascending profiles $\langle b\rangle$ is generated by uniformly sampling $p-1$ numbers in the interval [0,1] and sorting them in ascending order, for all criteria. Moreover, the voting weights $\langle w\rangle$ are generated by sampling $|\mathcal{N}|-1$ numbers in the interval [0,1], sort them, and use them as the cumulative sum of weights. The parameter λ is then randomly chosen with uniform probability in the interval]0.5,1[. Finally, we sample uniformly $|\mathbb{X}^*|$ tuples in $[0,1]^{\mathcal{N}}$, defining the performance table of the reference alternatives, and assign them to categories in $C^1 \prec \ldots \prec C^p$ according to the model $M^0 = \text{MR-Sort}_{\langle b\rangle,\langle w\rangle,\lambda}$ with the generated profiles, voting weights, and qualified majority threshold.

Solving the SAT problems:

We then translate the assignment into a binary satisfaction problem, described by sets of variables and clauses for both formulations as described in Sections 3 and 4. Each one of these binary satisfaction problems are written in a file, and passed to a command line SAT solver - CryptoMiniSat 5.0.1 [14]. The computing times t^{SAT^C} and t^{SAT^P} are recorded for both formulations (SAT^C and SAT^P). If the solver finds a solution, then it is converted into parameters ($\langle b^{SAT}\rangle, \langle T^{SAT}\rangle$) for a U-NCS model. Both formulations $M^{SAT^C} = \text{U-NCS}_{\langle b^{SAT^C}\rangle,\langle w^{SAT^C}\rangle,\lambda^{SAT^C}}$ and $M^{SAT^P} = \text{U-NCS}_{\langle b^{SAT^P}\rangle,\langle w^{SAT^P}\rangle,\lambda^{SAT^P}}$ yielded by the program are then validated against the input. As the ground truth M^0 used is an MR-Sort model (therefore a U-NCS model), we expect the solver to always find a solution. Moreover, the U-NCS model returned by the program should be "close" to the provided assignment.

Ability of the inferred models to restore the original one:

In order to appreciate how "close" a computed model $M^c \in \{M^{SAT^C}, M^{SAT^P}\}$ is to the ground truth from which the assignment examples were generated M^0, we proceed as follows: we sample a large set of n profiles in $\mathbb{X} = [0,1]^{\mathcal{N}}$ and compute the assignment of these profiles according to the original and computed MR-

Sort models (M^0 and M^c). On this basis, we compute $err - rate$ the proportion of "errors", i.e. tuples which are not assigned to the same category by both models. To obtain a reasonable sample for $\mathbb{X}$, we vary size of the sample of $\mathbb{X} = [0,1]^{\mathcal{N}}$ according to the number of criteria $|\mathcal{N}| : n = Max(Min(4^N, 3.10^5), 10^4)$.

For the second formulation, the values of variables do not define a single set of sufficient coalitions, and the formulation can return a set $\mathcal{T}$ of sufficient coalitions among compatible ones. As defined in Section 4, we can consider $\mathcal{T}_{min}$ and $\mathcal{T}_{max}$ the minimum and maximum compatible $\mathcal{T}$, and $\mathcal{T}_{rand}$ a randomly chosen compatible $\mathcal{T}$. In the first formulation, the variables define a unique set of sufficient coalitions.

5.2 Computing time

Figure 3 displays the time needed (logarithmic scale) by the second formulation (SAT^P) with 3 categories to compute M^{SAT^P}, versus the number of reference alternatives $|\mathbb{X}^*|$ and varying the number of criteria. A similar trend is observed for 2, 4 and 5 categories.

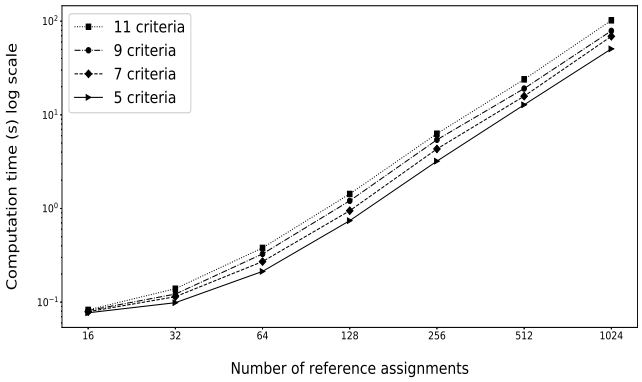

Figure 3. Computation time by size of the learning set (3 categories)

The observed linear trend in the results seem to exhibit a linear dependency between $log(t^{SAT^P})$ and $log(|\mathbb{X}^*|)$. The curves with a varying number of criteria, are almost parallel (with a slope close to 2), which seem to show that t^{SAT^P} is proportional to $|\mathbb{X}^*|^2$. The same observations in the plane (number of criteria * computation time) (not represented) lead to infer a law :

$$t^{SAT^P} \sim |\mathbb{X}^*|^2 * |\mathcal{N}|$$

For both formulations under scrutiny and the set of considered parameters governing the input, the computation time ranges from below the tenth of a second to a couple of minutes. Table 5 depicts the distribution of the computation time for the baseline situation (9 criteria, 3 categories, 128 reference assignments), the computation time of the second formulation appears about two times slower than the first formulation, on the other hand, the distribution of the computing time for each one of two formulations is very tight around its center.

In order to assess the influence of the parameters governing the size and complexity of the input, we explore situations differing from the baseline on a single parameter:

- The number of reference assignments $\mathbb{X}^*$: Figure 4 indicates that the distribution of the computing time for the two SAT-formulations remains tightly grouped around its central value. It also shows that this value steadily increases with the number of reference assignments. For the first formulation, $log(t^{SAT^C})$ is seemingly linearly ($O(|\mathbb{X}^*|)$) dependent on $log(|\mathbb{X}^*|)$, on the other hand, the computation time of the second formulation quadratically ($O(|\mathbb{X}^*|^2)$) increases with the number of reference assignments.

	Median	**1^{st} quartile**	**2^{nd} quartile**	**Min**	**Max**
$\boldsymbol{SAT^C}$	$10.95s$	$8.84s$	$12.95s$	$5.33s$	$17.04s$
$\boldsymbol{SAT^P}$	$5.79s$	$4.255s$	$6.759s$	$1.879s$	$8.138s$

Table 5. Computation time for both formulations in the baseline configuration: 9 criteria, 3 categories, 128 reference alternatives.

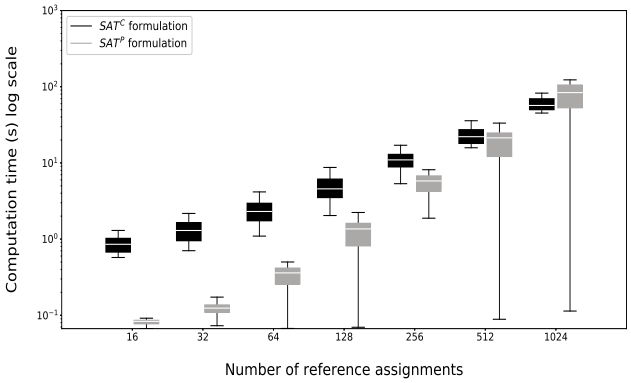

Figure 4. Computation time for both formulations by number of reference assignments (3 categories, 9 criteria)

- The number of criteria $\mathcal{N}$: Figure 5 indicates that the distribution of the computing time for both formulations remains tightly grouped around its central value.The computing time of SAT^C increases exponentially with the number of criteria ($O(2^{\mathcal{N}})$), while computing time of SAT^P increases linearly ($O(\mathcal{N})$).

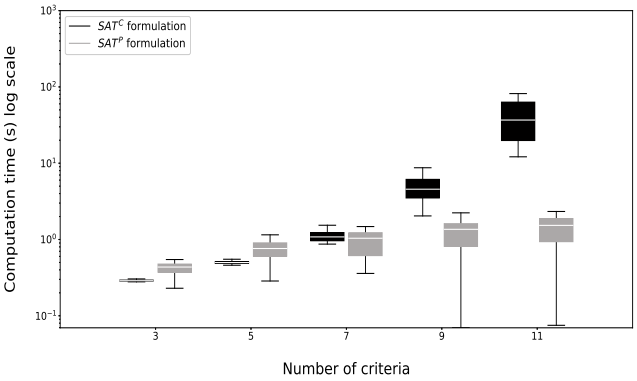

Figure 5. Computation time for both formulations by number of criteria (3 categories, 128 reference assignments)

- The number of categories p: for both SAT formulations, the experimental results display a linear dependence with a low slope (not represented) between the distribution of the computation time and the number of categories.

5.3 Results on the ability of the inferred model to restore the original one

The second formulation returns an acceptable set of sufficient coalitions. To identify the upset that best restores the basic MR-Sort model M^0, it will require to study the three following situations: $\mathcal{T} = \mathcal{T}_{min}$, $\mathcal{T} = \mathcal{T}_{rand}$ and $\mathcal{T} = \mathcal{T}_{max}$. T-Student tests ($p = 0.05$) show that the restoration rate when $\mathcal{T} = \mathcal{T}_{min}$ is always at least as good as the other two variants regardless of the number of criteria, alternatives and categories. For comparison with the first formulation, we tend to advocate the best, so we will consider only $\mathcal{T} = \mathcal{T}_{min}$ (see Table 6).

	Median	1^{st} quartile	2^{nd} quartile	Min	Max
$\mathcal{T} = \mathcal{T}_{min}$	0.888	0.828	0.937	0.669	0.995
$\mathcal{T} = \mathcal{T}_{rand}$	0.869	0.827	0.927	0.679	0.995
$\mathcal{T} = \mathcal{T}_{max}$	0.864	0.811	0.925	0.651	0.995

Table 6. Restoration rate for the second formulation in the baseline configuration (3 categories, 9 criteria, 128 Reference assignments)

Table 7 depicts the distribution of the proportion of correct assignments (as compared to the ground truth) for the baseline situation (9 criteria, 3 categories, 128 reference assignments). The proportion of correct assignments of both formulations is almost the same with a slight difference on the median.

	Median	1^{st} quartile	2^{nd} quartile	Min	Max
SAT^C	0.858	0.767	0.939	0.656	1
SAT^P	0.849	0.780	0.943	0.606	1

Table 7. Restoration rate for both formulations in the baseline configuration (3 Categories, 9 Criteria, 128 Reference assignments)

Figures 6, 7 and 8 present the variations of the alignment of the models yielded by both formulations with the ground truth with respect to the number of reference assignments of criteria or of categories, respectively. The experimental results display a tendency towards a degradation of this alignment as the number of criteria or the number of categories increases. Conversely, as expected, increasing the number of reference assignments noticeably enhances the restoration rate. The two formulations seem to behave in a similar manner with respect to the modification of these parameters.

5.4 Discussion

In this section, we discuss the influence of the input parameters (in particular number of criteria and size of the learning set) on on the computing time and ability to generalize.

On the one hand, the number of reference assignments impacts linearly the computation time of SAT^C and quadratically the computation time of SAT^P. On the other hand, SAT^C depends exponentially on the number of criteria, and this dependence remains linear for SAT^P.

For a fixed number of criteria, if we increase the number of reference assignments, SAT^C becomes faster than SAT^P starting from a threshold. These thresholds are calculated statistically using T-Student tests on both distributions. The results of our statistical tests are resumed in Table 8.

Table 8 represents the approximate thresholds from which SAT^C becomes faster than SAT^P for the corresponding number of criteria and categories. An example of 3 categories and 7 criteria is presented in Figure 9. For a number of reference assignments exceeding 128, SAT^C becomes faster.

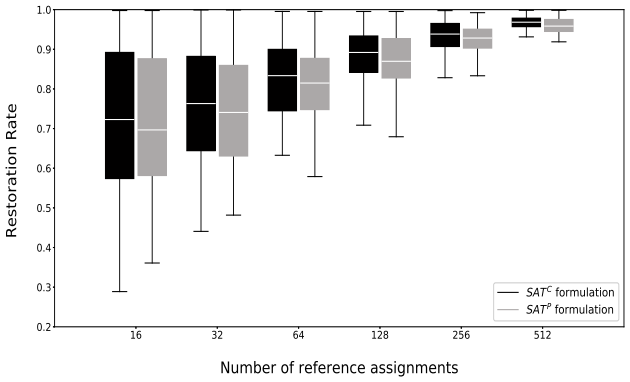

Figure 6. Restoration rate for both formulations by number of reference assignments (3 categories, 9 criteria)

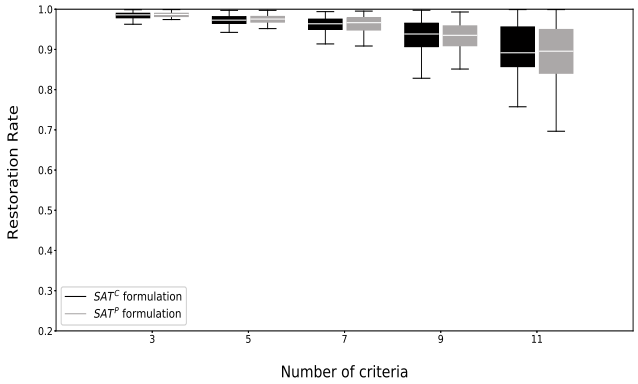

Figure 7. Restoration rate for both formulations by number of criteria (3 categories, 128 reference assignments)

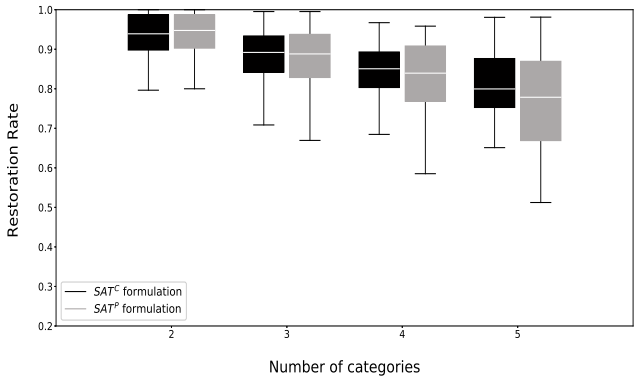

Figure 8. Restoration rate for both formulations by number of categories (9 criteria, 128 reference assignments)

	2 categ.	3 categ.	4 categ.	5 categ.
3 crit.	16	16	32	128
5 crit.	64	64	64	$]64, 128[$
7 crit.	128	$]128, 256[$	$]128, 256[$	256
9 crit.	$]512, 1024[$	$]512, 1024[$	$]512, 1024[$	$]512, 1024[$
11 crit.	>1024	>1024	>1024	>1024

Table 8. The approximate thresholds from which SAT^C becomes faster than SAT^P

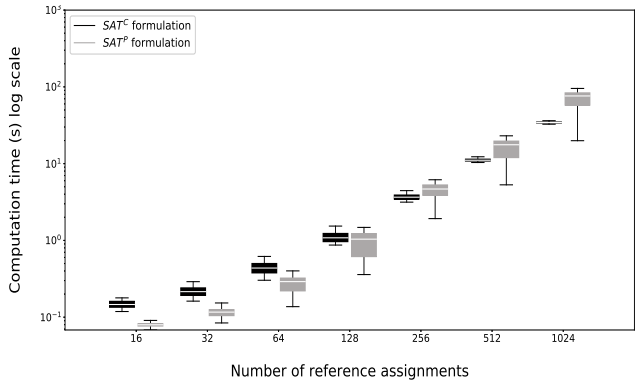

Figure 9. Computation time for both formulations by number of reference assignments (3 categories, 7 criteria)

For more than 5 criteria, and less than $\sim$150 examples, SAT^P is faster than SAT^C and the generalization is equivalent for both formulations. Above a certain threshold on the number of reference assignments, SAT^P is penalized by the fact that the number of clauses increase quadratically with the number of reference assignments (see Table 8). For more than 7 criteria, SAT^P is faster than SAT^C since the large number of criteria penalizes SAT^C by an exponential component ($t^{SAT^C} \sim |\mathbb{X}^*| * 2^{|\mathcal{N}|}$) and to properly calibrate the model, a large number of assignment examples is required.

6 Conclusion

In this paper, we consider the noncompensatory sorting model and evaluate the comparative performances of two alternative SAT formulations to infer the parameters of this sorting model from a learning set provided by the decision maker.

The results do not show significant differences between formulations in terms of generalization. Computation time of the two formulations evolves depending on the number of reference alternatives and the number of criteria; the second formulation performs better when the number of criteria increases, while it is the contrary when the number of reference alternatives increases.

However, for a real world decision problems (with more than 5 criteria and more than 100 references alternatives), the second formulation seems better as it is faster with an equivalent generalization.

This work opens avenue for further research. In particular, these formulations do not account for noisy input. Extending these formulations to the case where the set of assignment example is not fully compatible with NCS.

REFERENCES

[1] K. Belahcène, *algorithms Explaining multicriteria recommendations*, Ph.D. dissertation, Université Paris Saclay, CentraleSupelec, 2018.

[2] K. Belahcène, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane, 'An efficient SAT formulation for learning multiple criteria noncompensatory sorting rules from examples', *Computers & OR*, **97**, 58–71, (2018).

[3] K. Belahcène, C. Labreuche, N. Maudet, V. Mousseau, W. Ouerdane, and Y. Chevaleyre, 'Accountable approval sorting', in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*, (2018).

[4] K. Belahcene, V. Mousseau, M. Pirlot, and O. Sobrie, 'Preference elicitation and learning in a multiple criteria decision aid perspective', Technical report, Laboratoire Gnie Industriel, Ecole Centrale Paris, (February 2017). Research report 2017-02.

[5] D. Bouyssou and T. Marchant, 'An axiomatic approach to noncompensatory sorting methods in mcdm, i: The case of two categories', *EJOR*, **178**(1), 217–245, (2007).

[6] D. Bouyssou and T. Marchant, 'An axiomatic approach to noncompensatory sorting methods in mcdm, ii: More than two categories', *EJOR*, **178**(1), 246–276, (2007).

[7] D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs, and P. Vincke, *Evaluation and decision models with multiple criteria: Stepping stones for the analyst*, International Series in Operations Research and Management Science, Volume 86, Springer, Boston, 2006.

[8] M. Doumpos and C. Zopounidis, *Multicriteria Decision Aid Classification Methods*, Springer, 2002.

[9] J. Figueira, V. Mousseau, and B.phd Roy, 'Electre methods', in *Multiple criteria decision analysis: State of the art surveys*, 133–153, Springer, (2005).

[10] A. Leroy, V. Mousseau, and M. Pirlot, 'Learning the parameters of a multiple criteria sorting method', in *Algorithmic Decision Theory*, eds., R. Brafman, F. Roberts, and A. Tsoukiàs, volume 6992 of *Lecture Notes in Artificial Intelligence*, pp. 219–233, (2011). 2nd International Conference on Algorithmic Decision Theory, ADT 2011, Piscataway, NJ, USA.

[11] V. Mousseau, R. Słowiński, and P. Zielniewicz, 'A user-oriented implementation of the electre-tri method integrating preference elicitation support', *Computers & Operations Research*, **27**(7), 757–777, (2000).

[12] O. Sobrie, V. Mousseau, and M. Pirlot, 'Learning a majority rule model from large sets of assignment examples', in *Algorithmic Decision Theory*, eds., P. Perny, M. Pirlot, and A. Tsoukiás, volume 8176 of *Lecture Notes in Artificial Intelligence*, pp. 336–350. Springer, (2013).

[13] O. Sobrie, V. Mousseau, and M. Pirlot, 'Learning the parameters of a non compensatory sorting model', in *Algorithmic Decision Theory*, ed., T. Walsh, volume 9346 of *Lecture Notes in Artificial Intelligence*, pp. 153–170, Lexington, KY, USA, (2015). Springer.

[14] M. Soos, 'The cryptominisat 5 set of solvers at sat competition 2016', in *Proceedings of SAT Competition 2016: Solver and Benchmark Descriptions, volume B-2016-1 of Department of Computer Science Series of Publications B, University of Helsinki*, (2016).

D

Learning parameters of RMP with a SAT formulation

Ranking with Multiple reference Points: Efficient Elicitation and Learning Procedures

Khaled Belahcène[1], **Vincent Mousseau**[1], **Wassila Ouerdane**[1], **Marc Pirlot**[2] and **Olivier Sobrie**[2]

Abstract. We consider the multicriteria ranking problem, and specifically a ranking procedure based on reference points recently proposed in the literature, named Ranking with Multiple reference Points (RMP) [25, 8]. Implementing the RMP method in a real world decision problem requires to elicit the model preference parameters. This can be done indirectly by inferring the parameters from stated preferences, as in [21, 22, 12].

Learning an RMP model from stated preferences proves however to be computationally extremely costly, and can hardly be put in practice using state of the art algorithms. In this paper, we propose a Boolean satisfiability formulation of the inference of an RMP model from a set of pairwise comparisons which is much faster than the existing algorithms.

1 Introduction

The multiple criteria ranking problem consists in computing a preorder on a finite set of alternatives $\mathcal{A}$ when these alternatives are evaluated on multiple criteria. Many ranking methods have been proposed in the literature to tackle this problem. Among ranking methods, the so called *outranking* methods (see e.g., [13, 9]) proceed by comparing alternatives on each criterion, then aggregate these preference relations relative to criteria into a ranking. Actually, with these methods, a ranking is not obtained directly. The preference relations on each criterion are first aggregated into an outranking relation. This is done for each pair of alternatives by considering only the preferences between these alternatives on all criteria, without taking into account the other alternatives. In such a way the independence of irrelevant alternatives (IIA) property of the well known Arrow's impossibility theorem [1] is satisfied. The drawback is that the outranking relation is not transitive in general due to the possible presence of Condorcet cycles [10]. In order to obtain a ranking, a further step, called *exploitation* is applied to the outranking relation. Transitivity is obtained at the cost of loosing the IIA property (which is an unavoidable consequence of Arrow's theorem).

However, outranking methods are well-suited for ranking problems involving qualitative criteria, as they only consider the ordinal aspect of evaluation (as opposed to a cardinal aspect which requires assessing trade-offs between differences of evaluations). A recently proposed outranking based ranking method [25, 8], Ranking with Multiple reference Points (RMP), keeps the specificity of considering ordinal data while fulfilling the IIA property. This statement apparently contradicts Arrow's theorem. Actually, this is not the case, due to the introduction of an additional ingredient, namely the reference points.

[1] LGI, CentraleSupélec, Université Paris-Saclay, Gif-Sur-Yvette, France
[2] MATHRO, Faculté Polytechnique, Université de Mons, Belgium

Respecting the IIA principle is particularly important when learning ranking models from data (e.g., pairwise comparisons). In particular, when the comparisons involve real alternatives, learning a ranking model from comparisons can lead to a situation where: (*i*) the decision maker states that a is better than b ($a \succ b$), (*ii*) a ranking model $\mathcal{M}$ is computed from a learning set (including $a \succ b$), but (*iii*) when applying $\mathcal{M}$ to the set of alternatives, b is ranked better than a.

To the best of our knowledge, RMP is the only outranking based method which fulfils the IIA property; this ranking method is therefore well suited to be put in practice using learning algorithms that learn an RMP model from a set of pairwise comparisons. In our paper, we propose efficient tools to learn RMP models from data.

The paper is organized as follows. Section 2 introduces the RMP method. In Section 3, we present how to implement the RMP method in practice using algorithms that learn an RMP from pairwise comparisons provided by the Decision Maker (DM). We propose, in Section 4, a standard sequence procedure to elicit an RMP model. Section 5 describes a new efficient algorithm that computes an RMP model from a learning set. This algorithm is based on a Boolean satisfiability formulation. We perform, in Section 6, an empirical analysis of our algorithm to assess its performance as compared to the existing literature. A final Section groups conclusions and further research directions.

2 Ranking with Multiple Points

2.1 Reference points in multicriteria decision aid

Kahneman and Tversky were the first to identify clearly the role of reference points in the formation of preferences in the context of risky [16] and riskless decisions [27]. Reference based preferences have since been studied (see [18, 19]) and multicriteria models using reference points have been proposed to sort alternatives into categories (see e.g. [6, 7]), and to rank alternatives ([25, 8]). In this paper we consider the RMP ranking method [25].

2.2 An introductory example

To introduce how the Ranking with Multiple Points (RMP) method proceeds, we consider a simple illustrative example in which a set of cars are to be ranked from the best to the worst. We consider three cars x, y and z evaluated on the following four criteria: Brakes ([0-10] scale), Road holding ([0-10] scale), Price (€), and Acceleration (seconds to accelerate from 0 to 100km/h). The first two criteria are to be maximized, the last two are to be minimized. The performances of cars are shown in Table 1.

The RMP ranking method makes use of preference parameters to specify the decision maker judgment: (*i*) a set of reference points,

	Brakes (Max, [0,10])	Road holding (Max, [0,10])	Price (Min, €)	Acceleration (Min, sec.)
x	9.5	9.5	11.7 K€	29.4 sec.
y	0.5	9.5	11.8 K€	27.9 sec.
z	5	9.5	15.9 K€	26.7 sec.
r^2	8	8	12.0 K€	28.0 sec.
r^1	2	4	18.0 K€	31.0 sec.

Table 1. Illustrative example

and (*ii*) an importance relation on criteria coalitions (in this example, all criteria are assumed equally important, and it is sufficient to count criteria in coalitions to compare them).

In our example, we use two reference points (which are vectors of evaluations), r^1 and r^2, such that r^2_j is better than r^1_j on each criterion j. These two reference points define three segments of performances on each criterion:

- better than r^2 (which can be interpreted as "*good*"),
- between r^1 and r^2 (which can be interpreted as "*intermediate or fair*"); and
- worse than r^1 (which can be interpreted as "*insufficient*").

The values of these points r^1 and r^2 on criteria are provided in Table 1. For instance, on the criterion "Brakes", any alternative evaluated 8 or above will be considered "good" (e.g., alternative x) and any alternative evaluated lower than 2 will be considered "insufficient" (e.g., alternative y). In other terms, the reference points allow to identify an ordered encoding for each criterion defined by 3 ordered intervals of performances (A, B and C) as illustrated in Figure 1, such that:

A performances above r^2 on each criterion are denoted as A (which can be interpreted as "*good*").
B performances between r^1 and r^2 on each criterion are denoted as B (which can be interpreted as "*intermediate or fair*").
C performances below r^1 on each criterion are denoted as C (which can be interpreted as "*insufficient*")

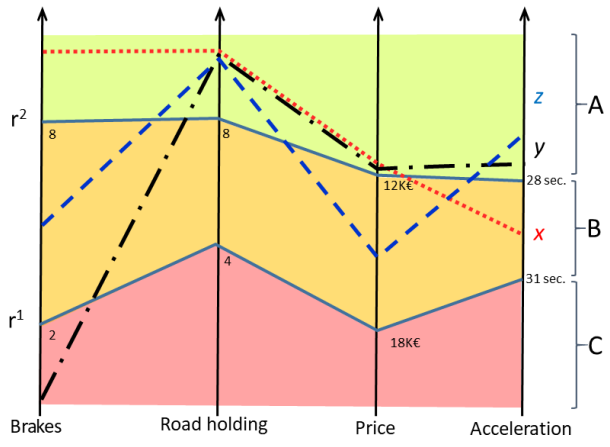

Figure 1. Graphical interpretation of Table 1

The RMP method ranks alternatives based on these ordered intervals of performances. Table 2 shows the results of the encoding for the 3 alternatives considered in our example. For instance, z is encoded B on criterion "Brakes" because z is worse than r^2 but better than r^1.

	Brakes	Road Holding	Price	Acceleration
x	A	A	A	B
y	C	A	A	A
z	B	A	B	A

Table 2. Results of the encoding procedure for the illustrative example

To compute a ranking, alternatives are not compared one to each other but compared to the reference points. Alternatives are compared to the first reference point r^1. Considering two alternatives a and b, a is preferred to b, noted $a \succ b$, if the coalition of criteria for which alternative a is evaluated A or B (i.e. better than r^1) is more important than the coalition of criteria for which alternative b is evaluated A or B (i.e., better than r^1). In this example, criteria are assumed equally important, so we just count the number of criteria. If a and b cannot be distinguished with respect to their comparison to r^1, then a and b are compared to r^2. If the number of criteria for which alternative a is evaluated A (i.e. better than r^2) is greater than the number of criteria for which alternative b is evaluated A (i.e., better than r^2), then a is preferred to b, otherwise a is indifferent to b. In our example, we thus have the following:

- Alternative x is better than y because x has evaluation A or B for all criteria, while y has evaluation A or B for only three criteria (x compares better to r^1 than y does).
- Alternative x is better than z because x and z are both evaluated A or B on all criteria (they compare equally to r^1), but x is evaluated A on three criteria while z is evaluated A only on two criteria (x compares better to r^2 than z does).
- Alternative z is better than y because z has evaluation A or B on all criteria while y has evaluation A or B on three criteria only (z compares better to r^1 than y does).

2.3 The RMP ranking method

We consider $\mathcal{A}$, a set of alternatives evaluated on n criteria. Let us denote $\mathcal{N} = \{1, 2, \ldots, i, \ldots, n\}$ the set of criteria indices, and a_i denotes the evaluation of alternative $a \in \mathcal{A}$ on criterion i (in what follows we will consider, without loss of generality, that preferences increase with the evaluation on each criterion, i.e., the greater the better). The RMP method is a method for ranking a finite set of alternatives evaluated on several criteria [25].

To rank alternatives, RMP compares alternatives to reference points, and then aggregates these comparisons into a final ranking. A dominance structure can be assumed on the set of reference points without loss of generality (for any RMP model using a set of reference points without any dominance structure, there exist an equivalent RMP model using a set of reference points with a dominance structure). RMP makes use of two types of preference parameters:

- $\mathcal{R} = \{r^1, r^2, \ldots, r^h, \ldots, r^m\}$, with $r^h = \{r^h_1, ..., r^h_i, ..., r^h_n\}$, where r^h_i denotes the evaluation of r^h on criterion i;
- an importance relation on criteria coalitions, $\trianglerighteq \subseteq \mathcal{P}(\mathcal{N})$, where $\triangleright$ and $\equiv$ represent the asymmetric and symmetric part of $\trianglerighteq$.

RMP proceeds through the following three steps:

1. compute $c(a, r^h) = \{i \in \mathcal{N} : a_i \geq r_i^h\}$, $a \in \mathcal{A}$, $h = 1, \ldots m$, the set of criteria on which alternative a is at least as good as the reference point r^h.
2. compare alternatives one to each other to define k preference relations $\succsim_{r^h}$ relative to each reference point such that $a \succsim_{r^h} b$ iff $c(a, r^h) \supseteq c(b, r^h)$. In other words, $a \succsim_{r^h} b$ holds when a compares better to r^h than b does. We denote $\succ_{r^h}$ ($\sim_{r^h}$, respectively) the asymmetric part of the relation $\succsim_{r^h}$ (the symmetric part of $\succsim_{r^h}$, respectively).
3. to rank two alternatives $a, b \in \mathcal{A}$, consider sequentially the relations $\succsim_{r^1}, \succsim_{r^2}, \ldots, \succsim_{r^k}$; a is preferred to b if $a \succ_{r^1} b$, or if $a \sim_{r^1} b$ and $a \succ_{r^2} b$, or ... Hence, a and b are indifferent iff $a \sim_{r^h} b$, for all $h = 1 \ldots m$.

Rolland [25] proved that by proceeding in such a way, the computed preference relations on alternatives are guaranteed to be transitive. As mentioned earlier, a dominance structure on the set of reference points can be assumed without loss of generality.

3 Implementing the RMP ranking method

To implement the RMP method in a decision aiding study, an interaction with the DM is required, so as to integrate her preferences, hence set the values of the preference parameters involved in the RMP method. A basic approach called *direct elicitation* consists in interacting with the DM directly on the values of the preference parameters. However, such an approach is not recommended as the DM usually has no clear understanding of the semantics attached to the preference parameters. Moreover, it imposes a strong cognitive burden on the DM. Therefore, the literature frequently proposes an *indirect elicitation*, in which the DM expresses holistic preferences (i.e., pairwise comparisons of alternatives) from which the values of the preference parameters are inferred (see e.g. [5, 15, 24]).

Recent literature (see [21, 28]) proposed indirect elicitation procedures for the S-RMP method (a particular case of RMP in which the criteria importance relation is additively representable). The 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. Two algorithms were proposed:

- **MIP-based algorithm.** [28, 21] formulate the elicitation of a S-RMP model as a mixed linear optimization problem. In this optimization program, the variables are the parameters of the S-RMP method, and additional technical variables which enable to formulate the objective function and the constraints in a linear form. The aim is to minimize the Kemeny distance (see [17]) between the partial ranking provided by the Decision Maker (i.e. the comparisons) and the S-RMP ranking. The resolution of this optimization program provides a guarantee 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.

- **Metaheuristic algorithm.** Another algorithm to indirectly elicit an S-RMP model, from pairwise comparisons, was proposed by [22, 21]. 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 to obtain an S-RMP model which 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 above mentioned algorithms suffer however from limitations:

- both algorithms only consider an additive representation of criteria importance relation, which can be restrictive when interaction between criteria occur;
- the MIP based approach is not able to deal with datasets whose size correspond to real world decision problems (e.g. 10 criteria, 2 reference points and 50 comparisons);
- the heuristic approach is fast but is not always able to restore an S-RMP model compatible with a set of comparisons, whenever it exists.

To circumvent these limitations, two paths are possible:

- elicit an RMP model using a *model-based elicitation strategy* analogous to the one described in [3] for the NonCompensatory Sorting model [6, 7]. This approach permits to elicit the RMP parameters by asking the decision maker to make comparisons, and aims at building the shortest questionnaire. We propose in Section 4 such a procedure for RMP with one single preference point.
- Design an algorithm similar to the MIP approach that can handle real-world size datasets, as done for the NonCompensatory Sorting model [6, 7] to overcome computational issues of [20] using a Boolean satisfaction (SAT) formulation, see [2]. In this perspective, we propose, in Section 5, a SAT formulation which is computationally efficient.

4 A procedure to elicit an RMP model

In this section, we restrict ourselves to RMP with a single reference point, and we propose an elicitation procedure in which the DM answers a sequence of questions that will lead to a complete knowledge of the RMP parameters (the importance relation on coalitions, and the reference point).

This procedure is structured in two consecutive phases: in the first phase, the answers of the DM leads to define the $\trianglerighteq$ importance relation on criteria coalitions, the reference point being unknown, while the second phase aims at specifying the reference point. The possibility to identify, in the first phase, the $\trianglerighteq$ relation without knowledge on the reference point is based on the following remark. Consider the alternative x_A, with $A \subseteq \mathcal{N}$, having the best possible evaluation on criterion $i \in A$, and the worst possible evaluation on criterion $j \in \mathcal{N} \setminus A$. if $x_A \succ x_B$, then $A \trianglerighteq B$ and not $[B \trianglerighteq A]$ hold whatever the reference point. Hence, it is possible to determine $\trianglerighteq$ in the absence of knowledge on the reference point (note, however, that this is possible only with RMP models involving a single reference point).

The first phase of the algorithm aims at eliciting the $\trianglerighteq$ relation. Let us first recall that the relation $\trianglerighteq$ defined on $\mathcal{P}(\mathcal{N})$ is transitive and compatible with inclusion, i.e, for any pair of criteria coalitions $A, B \subseteq \mathcal{N}$, $B \subset A \Rightarrow A \trianglerighteq B$. Consider the "*minimal*" relation $\trianglerighteq^0$ containing pairs of coalitions corresponding to inclusion situations. Consider two coalitions that are not in $\trianglerighteq^0$. A positive answer to the question "*is x_A preferred to x_B*" will enrich $\trianglerighteq^0$ with the statement $A \trianglerighteq B$, and all transitive consequences ($A' \trianglerighteq B'$, for all A', B' such that $A \subseteq A'$ and $B' \subseteq B$).

Hence, the answer to the question "*Is x_A preferred to x_B*?" will enrich relation $\trianglerighteq$, and we can proceed so until $\trianglerighteq$ corresponds to a complete pre-order. In other words, $\trianglerighteq^0$ should be completed to reach a complete and transitive relation on the subsets of $\mathcal{N}$, in which case

the importance relation on coalitions is fully known. Obviously, the order by which questions are posed should be defined so as to minimize the total number of questions. This issue is not discussed in this paper.

The second phase of the algorithm aims at eliciting the reference point r given the elicited relation $\trianglerighteq$. In order to elicit r_i the evaluation of the reference point on criterion i, consider two coalitions A and B, such that $i \notin B$, $A \trianglerighteq B$ and not $[A \trianglerighteq B \cup \{i\}]$. By construction, we have $x_A \succsim x_B$, but not $x_A \succsim x_{B\cup\{i\}}$. Consider now the alternative $x_B^{k_i}$ having the same evaluations as x_B except on criterion i on which its evaluation is k_i. If $x_A \succsim x_B^{k_i}$ holds, then it means that $k_i < r_i$. From the preceding implication, we can design a dichotomous search to elicit r_i from questions of the type "*Is x_A preferred to $x_B^{k_i}$?*". Proceeding in this way for each criterion leads to elicit r. Note that r_i can also be elicited analogously considering two coalitions A and B, such that $i \notin A$, not $A \trianglerighteq B$ and $A \cup \{i\} \trianglerighteq B$.

5 Learning an RMP model from pairwise comparisons: a SAT formulation model

In this section, we propose a new procedure to check whether a set of pairwise comparisons can be represented by an RMP model with k reference points using a Boolean satisfiability (SAT) formulation.

5.1 Boolean satisfiability (SAT)

A Boolean satisfaction problem consists of a set of Boolean variables V and a logical proposition about these variables $f : \{0,1\}^V \to \{0,1\}$. A solution $v^\star$ is an assignment of the 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 in the variables or their negation $\forall c \in \mathcal{C}, \exists c^+, c^- \in \mathcal{P}(V) : c = \bigvee_{v\in c^+} v \vee \bigvee_{v\in c^-} \neg v$, so that a solution satisfies at least one condition (either positive or negative) of every clause.

The models presented hereafter make extensive use of clauses where there is only one non-negated variable (a subset of *Horn clauses*): $a \vee \neg b_1 \vee \cdots \vee \neg b_n$, which represent the logical implication $(b_1 \wedge \cdots \wedge b_n) \Rightarrow a$. It is known since Cook's theorem [11] that the Boolean satisfiability problem is NP-complete. Consequently, unless $P = NP$, we should not expect to solve generic SAT instances quicker than exponential time in the worst case. Nevertheless, efficient and scalable algorithms for SAT have been – and still are – developed, and are sometimes able to handle problem instances involving tens of thousands of variables and millions of clauses in a few seconds (see e.g. [23, 4]).

5.2 A SAT encoding of given comparisons in RMP

We consider a set $\mathcal{BC} = \bigcup_{j\in\mathcal{J}}\{p^j \succ n^j\}$ of binary comparisons provided by the DM, (p for "positive", n for "negative"). Below, we will use the following indices:

- $h \in \mathcal{H}$ is an index for reference points ordered by importance (i.e. to compare alternatives, we consider r^1, then r^2 if needed, etc.);
- $i \in \mathcal{N}$ is the index for criteria;
- $j \in \mathcal{J}$ is the index for comparisons in the learning set, composed of pairs $p^j \succ n^j$ (p for "positive", n for "negative"), where $p^j = (p_1^j, p_2^j, \ldots, p_n^j)$ and $n^j = (n_1^j, n_2^j, \ldots, n_n^j)$ are evaluation vectors;
- $k \in \mathbb{X}_i$ denotes values taken on criterion $i \in N$ (i.e. the evaluation scale on criterion i is $\mathbb{X}_i = \bigcup_{j\in\mathcal{J}}\{p_i^j, n_i^j\}$).

We introduce the following variables:

- $x_{i,h,k}$ take value 1 iff the value k is above the reference point r^h on criterion i ($k \geq r_i^h$);
- $y_{A,B}$ take value 1 iff the criteria coalition A is more important than coalition B;
- $z_{j,h}$ take value 1 iff criteria for which alternative p^j is above reference point r^h are at least as important as those for which alternative n^j is above r^h $\big(c(p^j, r^h) \trianglerighteq c(n^j, r^h)\big)$;
- $z'_{j,h}$ take value 1 iff criteria for which alternative n^j is above reference point r^h are at least as important as those for which alternative p^j is above r^h $\big(c(n^j, r^h) \trianglerighteq c(p^j, r^h)\big)$;
- $d_{h,h'}$ take value 1 iff the reference point r^h dominates reference point $r^{h'}$ ($r_i^h \geq r_i^{h'}, \forall i \in \mathcal{N}$);
- $s_{j,h}$ take value 1 iff alternative p^j is indifferent to alternative n^j with respect to all reference points $r^{h'}$, with $h' < h$, and p^j compares to reference point r^h at least as well as n^j does;

Definition 1 (SAT encoding for RMP)**.** *Consider $\mathcal{BC} = \{(p^j, n^j), j \in \mathcal{J}\}$ a set of binary comparisons ($p^j \succ n^j$). We define the Boolean function $\phi_{\mathcal{BC}}^{SAT}$ as the conjunction of clauses:*

- *For all criteria $i \in \mathcal{N}$, for all reference point r^h, for all pairs of values $k, k' \in \mathbb{X}_i$ such that $k < k'$:*

$$x_{i,h,k} \vee \neg x_{i,h,k'} \tag{1}$$

- *For all pairs of reference points $r^h, r^{h'}$ such that $h < h'$:*

$$d_{h,h'} \vee d_{h',h} \tag{2a}$$

- *For all criteria $i \in \mathcal{N}$, for value $k \in \mathbb{X}_i$, for all pairs of reference points $r^h, r^{h'}$ such that $h \neq h'$:*

$$x_{i,h',k} \vee \neg x_{i,h,k} \vee \neg d_{h,h'} \tag{2b}$$

- *For all pairs of coalitions $A, B \subseteq \mathcal{N}$:*

$$y_{A,B} \vee y_{B,A} \tag{3a}$$

- *For all pairs of coalitions $A, B \subseteq \mathcal{N}$ such that $A \subset B$:*

$$y_{B,A} \tag{3b}$$

- *For all pairs of coalitions $A, B, C \subseteq \mathcal{N}$:*

$$\neg y_{A,B} \vee \neg y_{B,C} \vee y_{A,C} \tag{3c}$$

- *For all pairs of coalitions $A, B \subseteq \mathcal{N}$, for all comparisons $j \in \mathcal{J}$, for all reference point $r^h, h \in \mathcal{H}$:*

$$\bigvee_{i\notin A} x_{i,h,p_i^j} \vee \bigvee_{i\in B} \neg x_{i,h,n_i^j} \vee y_{A,B} \vee \neg z_{j,h} \tag{4a}$$

- *For all pairs of coalitions $A, B \subseteq \mathcal{N}$, for all comparisons $j \in \mathcal{J}$, for all reference point $r^h, h \in \mathcal{H}$:*

$$\bigvee_{i\notin A} x_{i,h,n_i^j} \vee \bigvee_{i\in B} \neg x_{i,h,p_i^j} \vee y_{A,B} \vee \neg z'_{j,h} \tag{4b}$$

- *For all pairs of coalitions $A, B \subseteq \mathcal{N}$, for all comparisons $j \in \mathcal{J}$, for all reference point $r^h, h \in \mathcal{H}$:*

$$\bigvee_{i\in A} \neg x_{i,h,p_i^j} \vee \bigvee_{i\notin A} x_{i,h,p_i^j} \vee \bigvee_{i\in B} \neg x_{i,h,n_i^j} \vee \bigvee_{i\notin B} x_{i,h,n_i^j} \vee \neg y_{B,A} \vee z'_{j,h} \tag{4c}$$

- *For each comparison $j \in \mathcal{J}$:*

$$\bigvee_{h \in H} s_{j,h} \tag{4d}$$

- *For each comparison $j \in \mathcal{J}$, for all pairs of reference points $r^h, r^{h'}; h, h' \in \mathcal{H}$ such that $h < h'$:*

$$z_{j,h} \vee \neg s_{j,h'} \tag{5a}$$

- *For each comparison $j \in \mathcal{J}$, for all pairs of reference points $r^h, r^{h'}; h, h' \in \mathcal{H}$ such that $h < h'$:*

$$z'_{j,h} \vee \neg s_{j,h'} \tag{5b}$$

- *For all reference points $r^h, h \in \mathcal{H}$:*

$$\neg z'_{j,h} \vee \neg s_{j,h} \tag{5c}$$

In Definition 1, clauses (1) impose that evaluation scale is monotone with respect to reference points on each criterion $i \in \mathcal{N}$. It states that if evaluation k is above r^h on criterion i, then any evaluation $k' > k$ is also above r^h on criterion i (we assume without loss of generality that all criteria are to be maximized).

Clauses (2a-2b) impose a dominance structure on reference points. (2a) check that, for any pair of reference points, either r^h dominates $r^{h'}$ or the reverse. Clauses (2b) relate variables $x_{i,h,k}$ to variables $d_{h,h'}$ stating that if, on criterion i, evaluation k is above reference point r^h, but not above reference point $r^{h'}$, then $r^{h'}$ does not dominate r^h.

Clauses (3a-3c) guarantee that the importance relation $\trianglerighteq$ on criteria coalitions is consistently defined. Clauses (3a) ensure relation $\trianglerighteq$ to be complete, clauses (3b) ensure that $\trianglerighteq$ is compatible with inclusion, and clauses (3b) impose transitivity.

Clauses (4a-4d) guarantee that the pairs p_j, n_j compare such that $p_j \succ n_j$. Clauses (5a-5c) guarantee that, for any comparison $j \in \mathcal{J}$, when p_j and n_j are separated by reference point$r^{h'}$, p_j and n_j are indifferent with respect to all reference points r^h such that $h < h'$.

6 Numerical investigation of the SAT formulation

In this section, we study the performance of the formulation proposed in section 5.2, both intrinsic and comparative with respect to state-of-the-art techniques. We use a state-of-the-art SAT solver, in order to solve instances of the problem of learning an RMP model, given a set of pairwise comparisons. We begin by describing our experimental protocol, with some implementation details. Then, we provide the results of the experimental study concerning the computation time of our algorithm, and particularly the influence of the size of the learning set, and the number of criteria, as well as elements of comparison between existing approaches.

6.1 Experimental design

The algorithm we test takes as input a set of pairwise comparisons in which alternatives compared are described by a performance tuple on a set of criteria $\mathcal{N}$.

The performance is measured in practice, by solving actual instances of the problem and reporting the computation time required. This experimental study is run on an ordinary laptop running under linux, equipped with an i7-6600U CPU at i2.6 GHz and 20 GB of RAM.

Dataset generation.

In the scope of this paper, we only consider to use a carefully crafted, random dataset as an input. On the one hand, the algorithm we describe is not yet equipped with the capability to deal with noisy inputs, so we do not consider feeding it with actual preference data, such as the one found in preference learning benchmarks [14]. On the other hand, using totally random, unstructured, inputs makes no sense in the context of algorithmic decision. Hence, we use a decision model to generate it, and, in particular, a model compatible with the RMP model. Precisely, we use a S-RMP model for generating the learning set, a model that particularizes RMP by postulating the set of importance relation on criteria coalitions possess an additive structure (i.e., there is a set of weights $w_i, i \in \mathcal{N}$, with $w_i \geq 0, \forall i$ and $\sum_i w_i = 1$, such that $A \trianglerighteq B$ iff $\sum_{i \in A} w_i \geq \sum_{i \in B} w_i$). This choice ensures our SAT formulation should succeed in finding the parameters of a model compatible with all the pairwise comparisons in the input.

When generating a dataset, we consider the number of criteria $|\mathcal{N}|$, the number of comparisons $|\mathcal{J}|$, and the number of reference points m as experimental parameters.

We consider all criteria take continuous values in the interval $[0, 1]$. We generate a set of m reference points $\langle r \rangle$ by uniformly sampling m numbers in the interval $[0, 1]$ and sorting them in ascending order, for all criteria; we then randomly re-order the reference points. We generate criteria weights $\langle w \rangle$ by sampling $|\mathcal{N}| - 1$ numbers in the interval $[0, 1]$, sorting them, and using them as the cumulative sum of weights.

Finally, we sample uniformly pairs of tuples in $[0, 1]^{\mathcal{N}}$, defining the performance of two alternatives[3], compare these two alternatives with $\mathcal{M}^0 :=$S-RMP$_{m,\langle r \rangle, \langle w \rangle}$ and consequently determine which one is p^j and n^j, $j \in \mathcal{J}$.

Solving the SAT problem.

For a given number of criteria $|\mathcal{N}|$, a given number of reference points m, we check if a given set $\mathcal{BC}$ of binary comparisons can be represented by the RMP model, by solving the corresponding SAT formulation presented in §5.2, using the SAT solver CryptoMiniSAT 5.0.1 [26], winner of the incremental track at SAT Competition 2016 (http://baldur.iti.kit.edu/sat-competition-2016/). If the solver finds a solution, then it is converted into parameters $(\langle r^{\mathrm{SAT}} \rangle, \trianglerighteq^{\mathrm{SAT}})$ for an RMP model. The model $\mathcal{M}^{\mathrm{SAT}} =RMP_{\langle r^{\mathrm{SAT}} \rangle, \trianglerighteq^{\mathrm{SAT}}}$ yielded by the program is then validated against the input. As the ground truth $\mathcal{M}^0$ used to generate the binary comparisons is an S-RMP model (and therefore an RMP model), we expect the solver to always find a solution, and we expect the RMP model returned by the program to always succeed at restoring the provided comparisons.

Ability of the inferred models to restore the original one.

In order to appreciate how "close" a computed model $\mathcal{M}^{SAT}$ is to the ground truth $\mathcal{M}^0$ from which the comparisons were generated, we proceed as follows: we sample a set of 10000 pairs of tuples in $X = [0, 1]^{\mathcal{N}}$ and compute the comparisons of these pairs according to the original and computed RMP models ($\mathcal{M}^0$ and $\mathcal{M}^{SAT}$). On this basis, we compute $err - rate$ the proportion of "errors", *i.e.* pairs which do not compare in the same way by both models.

[3] Only pairs of tuples that are not in the dominance relation are kept.

6.2 Performance of the SAT formulation

We run the above described experimental protocol varying the various values of the parameters: (*i*) the number of criteria $|\mathcal{N}|$ is chosen among $\{3, 4, 5\}$, (*ii*) the number of comparisons $|\mathcal{BC}|$ is chosen among $\{100, 200, ..., 1000\}$, and (*iii*) the number of reference points m is chosen among $\{1, 2, 3\}$. For each value of the triplet of parameters, we sample 10 S-RMP models $\mathcal{M}^0$, and record the computation time (t) needed to provide a model $\mathcal{M}^{\text{SAT}}$.

6.2.1 *Results regarding computation time.*

Figures 2 and 3 show the average computing time required to infer the parameters of one RMP model when the number of examples, criteria and reference points vary. We see in Fig. 2 that the computing time seems to grow exponentially as a function of the number of criteria. Indeed, when the reference set contains 500 alternatives, the average computing times for 2 reference points and 3, 4 and 5 criteria are respectively equal to about 1.5 seconds, 15 seconds and 75 seconds. It is no surprise since the number of constraints in the SAT formulation evolves as well exponentially as the number of criteria grows. When we vary the number of reference points (Figure 3), we observe that the same phenomenon occurs. Indeed, for 500 pairwise comparisons in the learning set, the average computing time is about 20 seconds when the model has one reference point, it grows up to ± 60 seconds for two reference points and up to ± 250 seconds for 3 reference points. For an RMP model with a fixed number of criteria and reference points, we see both in Figures 2 and 3 that the computing time evolves linearly when the number of pairwise assignment increases. Again, this is no surprise since the number of constraints involved also tends to increase linearly.

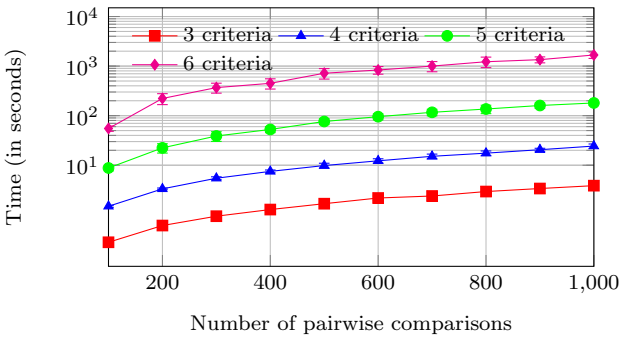

Figure 2. Computing time as a function of the number of pairwise comparisons for models involving 2 reference points and 3 to 6 criteria. Bars represent standard deviation.

6.2.2 *Results on the ability of the inferred model to restore the original one.*

To assess the ability of the SAT formulation to restore a model that is the closest to the original one, we sample a set of 10000 pairwise comparisons and we compute their relation of preference both with the original model ($\mathcal{M}^0$) and the one learned with the SAT solver ($\mathcal{M}^{\text{SAT}}$). Then we compute the proportion of binary comparisons that have the same preference relation with $\mathcal{M}^0$ and $\mathcal{M}^{\text{SAT}}$.

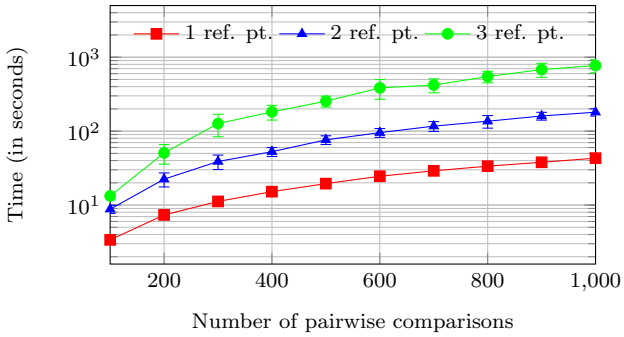

Figure 3. Computing time as a function of the number of pairwise comparisons for models with 5 criteria, and 1 to 3 reference points. Bars represent standard deviation.

In Figures 4 and 5, we observe that the average number of pairs of alternatives from the test set that have the same preference relation both with $\mathcal{M}^0$ and $\mathcal{M}^{\text{SAT}}$ increases as a function of the number of pairs in the learning set. When the number of criteria increases, the number of pairs required to restore the original model $\mathcal{M}^0$ increases. Figure 4 shows that with 100 alternatives, it is possible to restore on average more than 90 percent of the relations. With 6 criteria and a learning set of 100 pairs, less than 80 percent of the pairwise relations are restored. The same observation holds when the number of reference points increases (see Figure 5).

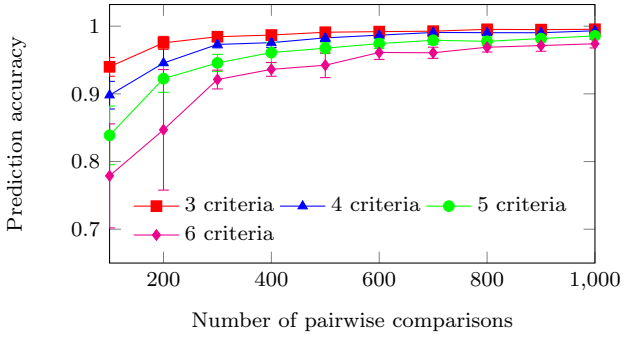

Figure 4. Average prediction accuracy as a function of the number of pairwise comparisons for models involving 2 reference points, and 3 to 6 criteria. Test set of 10000 pairwise comparisons. Bars represent standard deviation.

6.3 Discussion

Experimental results have shown that the algorithm was efficient for inferring an RMP model from large sets of binary comparisons. Indeed, the formulation is able to restore an RMP model composed of 3 reference points and 5 criteria from 500 binary comparisons in more or less 250 seconds. Furthermore, the algorithm performs well in generalisation. With barely 100 alternatives, the SAT formulation can learn an RMP model that predicts more than 70% of the binary relations obtained with a S-RMP model.

It should be highlighted that such a performance proves this formulation to be superior to existing algorithms. Indeed, MIP based

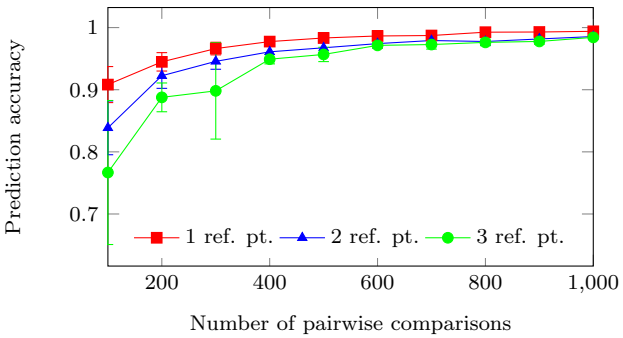

Figure 5. Average prediction accuracy as a function of the number of pairwise comparisons for models involving 5 criteria, and 1 to 3 reference points. Test set of 10000 pairwise comparisons. Bars represent standard deviation.

algorithms [28] are only able to handle a few dozens of pairwise comparisons which is insufficient to infer an RMP model with good generalization ability. Heuristic approaches [22, 21] can handle larger datasets, but are not able to systematically restore an RMP compatible input. A drawback of our approach is however its inability to easily handle noisy input.

7 Conclusion

In this paper, we describe a SAT formulation in order to learn an RMP model from a set of binary comparisons. Experimental results show that the algorithm is efficient enough to deal with large datasets and performs well in generalization. This formulation can be solved more efficiently than the MIP [28] and is more acurate than the heuristic approach [22, 21]. Our proposal is a step forward toward the possibility of eliciting an RMP model in an interactive process with the DM.

We see several research that should be pursued. The formulation presented in this paper can only deal with datasets that do not contain errors. A path to explore consists in finding a formulation that is able to handle errors, for instance by using a MAXSAT formulation. In this paper, the experiments have been done on artificial datasets. Another path to explore consists in using it with real datasets like in [12].

REFERENCES

[1] K.J. Arrow., *Social Choice and Individual Values*, Cowles Foundation Monographs; New York: Wiley, 1953.

[2] K. Belahcène, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane, 'An efficient SAT formulation for learning multiple criteria non-compensatory sorting rules from examples', *Computers & OR*, **97**, 58–71, (2018).

[3] K. Belahcène, V. Mousseau, M. Pirlot, and O. Sobrie, 'Preference elicitation and learning in a multiple criteria decision aid perspective', Technical report, Laboratoire Gnie Industriel, Ecole Centrale Paris, (February 2017). Research report 2017-02.

[4] A. Biere, M. Heule, H. van Maaren, and T. Walsh, *Handbook of Satisfiability*, Frontiers in Artificial Intelligence and Applications 185, IOS Press, 2009.

[5] G. Bous, P. Fortemps, F. Glineur, and M. Pirlot, 'ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements', *European Journal of Operational Research*, **206**(2), 435 – 444, (2010).

[6] D. Bouyssou and T. Marchant, 'An axiomatic approach to noncompensatory sorting methods in MCDM, I: The case of two categories', *EJOR*, **178**(1), 217–245, (2007).

[7] D. Bouyssou and T. Marchant, 'An axiomatic approach to noncompensatory sorting methods in MCDM, II: More than two categories', *EJOR*, **178**(1), 246–276, (2007).

[8] D. Bouyssou and T. Marchant, 'Multiattribute preference models with reference points', *European Journal of Operational Research*, **229**(2), 470 – 481, (2013).

[9] J-P. Brans and B. Mareschal, *Promethee Methods*, 163–186, Springer New York, New York, NY, 2005.

[10] M. Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris, 1785.

[11] S. Cook, 'The complexity of theorem proving procedures', in *Proceedings of the third annual ACM symposium on Theory of computing*, pp. 151–158, (1971).

[12] V. Ferretti, J. Liu, V. Mousseau, and W. Ouerdane, 'Reference-based ranking procedure for environmental decision making: Insights from an ex-post analysis', *Environmental Modelling & Software*, **99**, 11 – 24, (2018).

[13] J. Figueira, V. Mousseau, and B. Roy, *Electre Methods*, 133–153, Springer New York, New York, NY, 2005.

[14] J. Fürnkranz and E. Hüllermeier, *Preference Learning*, Springer-Verlag New York, Inc., 2011.

[15] E. Jacquet-Lagrèze and J. Siskos, 'Assessing a set of additive utility functions for multicriteria decision-making, the UTA method', *European Journal of Operational Research*, **10**(2), 151 – 164, (1982).

[16] D. Kahneman and A. Tversky, 'Prospect theory: An analysis of decision under risk', *Econometrica*, **47**(2), 263–291, (1979).

[17] J.G. Kemeny, 'Mathematics without numbers', *Daedalus*, **88**(4), 577–591, (1959).

[18] B. Koszegi and M. Rabin, 'A model of reference-dependent preferences', *The Quarterly Journal of Economics*, **121**(4), 1033 – 1065, (2006).

[19] C. Labreuche and M. Grabisch, 'Using multiple reference levels in multi-criteria decision aid: The generalized-additive independence model and the Choquet integral approaches', *European Journal of Operational Research*, **267**, 598 – 611, (2018).

[20] A. Leroy, V. Mousseau, and M. Pirlot, 'Learning the parameters of a multiple criteria sorting method', in *Algorithmic Decision Theory*, eds., R. Brafman, F. Roberts, and A. Tsoukiàs, volume 6992 of *Lecture Notes in Artificial Intelligence*, pp. 219–233, (2011). 2nd International Conference on Algorithmic Decision Theory, ADT 2011, Piscataway, NJ, USA.

[21] J. Liu, *Preference elicitation for multi-criteria ranking with multiple reference points*, Ph.D. dissertation, CentraleSupélec, Université Paris-Saclay, 2016.

[22] J. Liu, W. Ouerdane, and V. Mousseau, 'A metaheuristic approach for preference learning in multicriteria ranking based on reference points', in *Proceeding of the 2nd wokshop from multiple criteria Decision aid to Preference Learning (DA2PL)*, pp. 76–86, Chatenay-Malabry, France, (2014).

[23] M. Moskewicz, C. Madigan, Y. Zhao, L. Zhang, and S. Malik, 'Chaff: engineering an efficient SAT solver', in *Proceedings of the 38th annual Design Automation Conference (DAC '01). ACM, New York, NY, USA*, pp. 530–535, (2001).

[24] V. Mousseau and R. Słowiński, 'Inferring an ELECTRE TRI model from assignment examples', *Journal of global optimization*, **12**(2), 157–174, (1998).

[25] A. Rolland, 'Reference-based preferences aggregation procedures in multi-criteria decision making', *European Journal of Operational Research*, **225**(3), 479 – 486, (2013).

[26] M. Soos, 'The CryptoMiniSat 5 set of solvers at SAT Competition 2016', in *Proceedings of SAT Competition 2016: Solver and Benchmark Descriptions, volume B-2016-1 of Department of Computer Science Series of Publications B, University of Helsinki*, (2016).

[27] A. Tversky and D. Kahneman, 'Loss aversion in riskless choice. a reference-dependent model', *Quarterly Joournal of Economics*, **106**(4), 1039–1061, (1991).

[28] J. Zheng, A. Rolland, and V. Mousseau, 'Preference elicitation for a ranking method based on multiple reference profiles', Technical report, Laboratoire Génie Industriel, Ecole Centrale Paris, (August 2012). Research report 2012-05.

Titre : Explications pour l'agrégation des préférences — une contribution à l'aide à la décision responsable

Mots clés : Aide à la décision, Représentation des connaissances, Recherche opérationnelle, Intelligence artificielle

Résumé : Nous cherchons à équiper un processus d'aide à la décision d'outils permettant de répondre aux exigences de redevabilité. Un décideur fournit de l'information quant à ses préférences au sujet de la façon d'arbitrer entre des points de vue conflictuels. Un analyste, chargé d'éclairer la prise de décision, fait l'hypothèse d'un modèle de raisonnement, et l'ajuste aux informations fournies par le décideur. Nous faisons l'hypothèse d'un processus d'élicitation robuste, dont les recommandations sont déduites des éléments dialectiques. Nous nous sommes donc intéressés à la résolution d'un problème inverse concernant le modèle, ainsi qu'à la production d'explications, si possible correctes, complètes, facile à calculer et à comprendre. Nous avons considéré deux formes de représentation du raisonnement: l'une ayant trait à la comparaison de paires d'alternatives fondée sur un modèle de valeur additive, l'autre ayant trait au tri des alternatives dans des catégories ordonnées fondé sur un raisonnement non-compensatoire.

Title : Towards accountable decision aiding: explanations for the aggregation of preferences

Keywords : Decision Aiding, Knowledge Representation, Operations Research, Artificial Intelligence

Abstract : We consider providing a decision aiding process with tools aiming at complying to the demands of accountability. Decision makers, seeking support, provide preference information in the form of reference cases, that illustrates their views on the way of taking into account conflicting points of view. The analyst, who provides the support, assumes a generic representation of the reasoning with preferences, and fits the aggregation procedure to the preference information. We assume a robust elicitation process, where the recommendations stemming from the fitted procedure can be deduced from dialectical elements. Therefore, we are interested in solving an inverse problem concerning the model, and in deriving explanations, if possible sound, complete, easy to compute and to understand. We address two distinct forms of reasoning: one aimed at comparing pairs of alternatives with an additive value model, the other aimed at sorting alternatives into ordered categories with a noncompensatory model.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France