Explaining Fairness-Oriented Recommendations using Transfers and Transitive Arguments

Hénoïk Willot, Khaled Belahcene, Sébastien Destercke

Explaining Fairness-Oriented Recommendations using Transfers and Transitive Arguments

Hénoïk Willot[1] **Khaled Belahcene**[1] **Sébastien Destercke**[1]

Abstract. We explore the generation of explanations for preferences based on an ordered weighted average (OWA) model aiming to favour a balanced distribution of performances relative to the different points of view. In this paper we propose explanations, correct towards the model, based on transitivity with elementary arguments such as Pareto dominance, Pigou-Dalton transfers and the preferential information (PI) given by the decision maker. We propose several heuristic-based approaches allowing to compute such explanations, confirmed by experiments on their availability and their optimal length.

1 Introduction

Our aim is to propose explanation tools for recommendations based on preference aggregation. The need for decision-theoretic recommender systems–tools helping decision makers to formalize and support their judgment in a principled manner–has in turn given rise to a need for tools allowing people–be they actors of the decision process, or third parties impacted by it–to understand, scrutinize, validate or contradict the functioning of such recommender systems.

Preference aggregation is the process of merging comparative judgments expressed from various points of view into a single ranking. Points of view can represent various aspects of a situation (such as in multiple criteria decision aiding–MCDA), be expressed by various agents, or several possible worlds when modelling uncertainty. In turn, the aggregated judgment can be used as a basis for decision, supporting tasks such as choosing the best alternative, comparing them, or sorting them into ordered categories [3].

The MCDA literature usually distinguishes three approaches to aggregation: "aggregate then compare", where judgments are normatively described as complete preorders and described numerically with a score–the higher, the better–and the aggregator is a multi-attribute utility function [11]; "compare then aggregate", where the aggregated judgment is represented by an outranking relation constructed from the preference profiles [17]; and models based on logic. Following the numeric approach, it is customary to decide on several high-level features of the aggregator–either technical, such as possessing an additive form [12] , or decision-theoretic, such as being compatible to Pareto-dominance, satisfying anonymity or idempotence, etc. Usually, these requirements are chosen so as to define, either directly or indirectly via a representation theorem, a parametric family of aggregators. When the decision task requires to be able to compare any two alternatives, the usual approach, called preference elicitation, is to select a specific, precise value of the preference parameter. It is common to use indirect elicitation techniques, where the aggregator is fitted to preference information (PI) given by the decision maker in the form of comparative statements about alternatives, as opposed to statements concerning the parameters [10]. Full elicitation is not mandatory, though: skeptical recommendations can be derived considering the whole set of aggregators of the family that are compatible to the PI.

In the context of MCDA, Belahcene et al. have recently shown that, for the class of additive aggregators, it is possible to provide structured explanations where elementary arguments are organized according to a specific scheme. In [1], an explanation of a comparative statement is a decomposition into elementary swaps linked together by transitivity. In [2], an explanation is a decomposition into preference statements committed by the decision maker, assembled together by a high-order cancellation property.

When points of view are assessed on the same scale, the Choquet integral is a convenient class of aggregators, offering a good mixture of expressiveness, interpretabilty and computational tractability [7]. We focus on the subclass of *anonymous* aggregators, where the respective identities of the points of view play no role into their aggregation. Evaluations can be permuted, and the importance of a given score is related to its rank in the ordering of scores. These aggregators are thus named *ordered weighted average* (OWAs). Introduced in MCDA by Yager [19], they form a family of function parameterized by a tuple of weights, one per criteria, similarly to the weighted sum, and encompass the minimum, maximum, median and mean operators as particular cases. Moreover, by imposing the weights to be non-increasing w.r.t. the rank, it is possible to favour balanced scores over imbalanced ones, thus representing a sense of fairness. The explanatory engine described in [1] relies on swaps between criteria and is inspired by the notion of *even swaps* [9]. In the field of welfare economy, many methods and criteria are used to rank sets of incomes depending on the distribution of wealth among agents. The Pigou-Dalton principle [18] provides a similar notion of acceptable transfers in this context: the inequality between agents is reduced when a rich agent gives a small portion ϵ of its wealth to a poorer agent.

Our contribution is the definition of structured explanations for recommendations based on fairness-oriented OWAs. We propose to arrange comparative statements based on Pareto dominance, Pigou-Dalton transfers, and PI into a transitive structure. We begin by introducing the OWA operator, the Pigou-Dalton Principle and other definitions we will use in Section 2. In Section 3, we deal with the case where preference and explanations can be constructed without relying on preference information, and propose a heuristic to compute short explanations. In section 4, we address the case where preference is inferred from PI, and propose to find an additive decomposition of a comparative statement as an intermediate step towards finding a transitive explanation. Finally we will run an example combining

[1] University of Technologie of Compiègne,
email : {henoik.willot, khaled.belahcene, sebastien.destercke}@hds.utc.fr

both in Section 5 and give some insight about the performances of the method in Section 6.

2 Preliminaries

2.1 Ordered Weighted Averages

In this section we will introduce the decision problems and the formulation of the OWA operator. The MCDA problem we consider is ranking alternatives over a set of n criteria $N = \{1, \dots n\}$, defined on the same domain $\mathcal{X}$, which can be $[0, 1]$ or $\mathbb{R}$. The model of preference should then create an order over the candidates represented by their vectors $x \in \mathcal{X}^n$.

Definition 1 (Reordering function)**.** *We define the reordering function $^\uparrow$ as the permutation function over $\mathcal{X}^n$ s.t. $x \mapsto x^\uparrow$ with $x_1^\uparrow \leq x_2^\uparrow \leq \dots \leq x_n^\uparrow$.*
We denote by $\mathcal{X}^{n\uparrow}$ the domain of such vectors $x^\uparrow$.

Definition 2 (Ordered Weighted Average [19])**.** *The OWA operator is a function $\mathcal{X}^n \to \mathbb{R}^+$ defined by a vector of weights $w \in \mathcal{W}$ s.t. $\forall\, i\ w_i \in [0, 1]$ and $\sum_{i=1}^n w_i = 1$:*

$$OWA_w(x) = \sum_{i=1}^{n} w_i x_i^\uparrow$$

Even if it is presented as a weighted sum, thanks to the reordering function $^\uparrow$ the weights are assigned to the rank of the criteria, allowing to represent non linear preference operators, such as the min operator $w_{min} = (1, 0, \dots, 0)$, the max operator $w_{max} = (0, \dots, 0, 1)$, and in general any quantile as well as the arithmetic mean operator $w_{mean} = (\frac{1}{n}, \dots, \frac{1}{n})$.

Definition 3 (Fairness-oriented OWA (FOWA))**.** *Also called OWA operator with decreasing weights [6] or totally or-like OWA operator[16], an OWA operator is fairness-oriented if its weight vector w also satisfy $w_1 \geq w_2 \geq \dots \geq w_n$. We note by $\mathcal{W}^\searrow \subset \mathcal{W}$ the domain of such weights.*

A FOWA represents preferences oriented toward equity because, with a higher emphasis put on the smaller values, it gives higher scores to vectors with balanced values than to vectors where modalities are concentrated on a small subset of criteria. We can note that an increase in a small variable has a bigger impact on the aggregated value than the same increase in a big variable, *i.e.* $\forall\, i < j\ OWA_w(x^\uparrow + k \times e_i) \geq OWA_w(x^\uparrow + k \times e_j)$, with e_l the vector that is one for its l^{th} element and zero everywhere else.

Remark 1. *A FOWA with very unbalanced weights, i.e. tending toward the min operator (1 on the first criteria and 0 on the others), is more fairness oriented than a balanced operator with $\frac{1}{n}$ on each criteria, i.e. tending toward the arithmetic mean, even though its behaviour tends to balance the vectors of candidates.*

Definition 4 (Ranking relation)**.** *We define the ranking relation $\succeq_w$ induced by the operator OWA_w by :*

$$a \succeq_w b \iff OWA_w(a) \geq OWA_w(b)$$

Example 1. *The Decision Maker is asked to rank students over their results (scaled between 0 and 1) in their 3 main courses {"Physics", "Biology", "Maths"}. She prefers students who are balanced between the 3 courses, and with an analyst the OWA with the following vector of weights has been designed : $w = (0.6, 0.3, 0.1)$.*

If we consider three students a, b and c such that $a = (0.7, 1, 0.5)$, $b = (0.7, 0.7, 0.7)$ and $c = (1, 0.7, 0.8)$. Their scores defined by the OWA w will be computed on their reordered vectors $a^\uparrow = (0.5, 0.7, 1)$, $b^\uparrow = (0.7, 0.7, 0.7)$ and $c^\uparrow = (0.7, 0.8, 1)$, and are :

$$OWA_w(a) = 0.5 \times 0.6 + 0.7 \times 0.3 + 1 \times 0.1 = 0.61$$
$$OWA_w(b) = 0.7 \times 0.6 + 0.7 \times 0.3 + 0.7 \times 0.1 = 0.7$$
$$OWA_w(c) = 0.7 \times 0.6 + 0.8 \times 0.3 + 1 \times 0.1 = 0.76$$

Therefore we obtain the preferences $c \succeq_w b \succeq_w a$.

In this example, we do not exactly know why the weight vector was $w = (0.6, 0.3, 0.1)$. The issue with determining a specific set of weights is that it produces a total preorder (with possible ties) and may produce knowledge that the DM is not aware of and could potentially disagree with. Furthermore, obtaining precise values is cognitively demanding and require strong efforts. To circumvent this problem of finding the right set of weights, we can robustify our model using a set of models [14]. The set of OWA is defined as the set which respects the information obtained from the DM, her preferential information (PI), through an interactive process. In our case of study, the information collected is of the shape of m preference statements $a^j \succeq_{PI} b^j, j \in M = \{1, \dots, m\}$, with a^j, b^j alternatives.

Definition 5 (Robust OWA)**.** *We define a robust OWA operator the set $W_\mathcal{P} \subseteq \mathcal{W}$ of OWA weights :*

$$W_\mathcal{P} = \{w \in \mathcal{W} \ :\ \forall j \in M\ a^j \succeq_w b^j\}$$

And we note $W_\mathcal{P}^\searrow = \mathcal{W}^\searrow \cap W_\mathcal{P}$

As we now have a set of models instead of a single vector of weights, we have to adapt our process for producing preferences. We can define two relations of preferences from the set $W_\mathcal{P}$, a necessary and a possible preference relations [5]. In this paper we only focus on the necessary preference.

Definition 6 (Necessary preference)**.** *We define the necessary preference $\mathcal{N}_{PI}^{OWA^\searrow}$ of a robust FOWA with respect to preference information $\mathcal{P}$ as :*

$$a \mathcal{N}_\mathcal{P}^{OWA^\searrow} b \Leftrightarrow \forall w \in W_\mathcal{P}^\searrow,\ a \succeq_w b$$

In order to compute the set $W_\mathcal{P}$ and to reason with the necessary preference relation $\mathcal{N}_\mathcal{P}^{OWA^\searrow}$, we can adapt the GRIP method [4] that allows to decide whether a pair of alternatives belongs to the necessary preference relation, given some PI, for the additive value model, by solving a linear program[2]. In fact, to represent OWA operators, we only need to feed the method with vectors already reordered by $^\uparrow$, and for the representation of FOWA operators we need to add $n - 1$ linear constraints $w_i \geq w_{i+1}$ [3], $i \in \{1, \dots, n-1\}$ to constraint the weights of the additive value function to be decreasing. More details about the linear program formulation are given in Appendix A.

2.2 Pigou-Dalton Principle and Dominance relations

In this section we will first connect the OWA aggregators to the Pigou-Dalton principle. To do so, we will introduce the dominance

[2] Such a LP formulation could already be found in [8], but we opt to use the more streamlined formalism of GRIP.

[3] to follow the GRIP methods notations, the constraints are $u_i(\beta_i) - u_{i+1}(\beta_{i+1}) \geq 0$

relations and the Pigou-Dalton principle that we will use in our explanation engine and its potential use cases.

The Pigou-Dalton Principle was first introduced in welfare economic problem, where the components of the vector to compare are the incomes of economical agents, ranked from bottom to top [18]. In this context, the Pigou-Dalton Principle defines a relation $\succeq_{PDP}$ between two distributions over n agents.

Definition 7 (Pigou-Dalton Principle)**.** *Let x be the income vector of n agents such that $x = (x_1, \dots, x_n)$,*
$x_1 \leq x_2 \leq \dots \leq x_n$.
The vector x' is favoured to x by the Pigou-Dalton Principle, noted $x \preceq_{PDP} x'$ if there exists points of view $i, j \in N$, $i < j$ and quantity $\epsilon > 0$ s.t. :

$$\begin{cases} \forall k \in N,\ k \neq i,\ k \neq j,\ x'_k = x_k; \\ x'_i = x_i + \epsilon \ \leq x_{i+1};\ \textit{and} \\ x'_j = x_j - \epsilon \geq x_{j-1}. \end{cases}$$

When these conditions are met, we also use the notation $x \underset{PDP}{\overset{j \to i}{\preceq}} x'$ to account for the witnesses i and j.

The Pigou-Dalton Principle therefore favours vectors of incomes where a "rich" agent j gives a positive portion ϵ of its wealth to a "poorer" agent i in order to reduce the inequality. We also add explicitly another constraint on ϵ which is not always clear in the literature : the order in the distribution is preserved, $x_j - \epsilon \geq x_{j-1}$ and $x_i + \epsilon \leq x_{i+1}$. This principle of equity is respected by FOWA operators as we discussed earlier after definition 3.

We will now introduce two preorders relations, compatible with FOWA operators, which will be used by our explanation engine: the Pareto and Lorenz dominance, denoted respectively as $\succeq_P$ and $\succeq_L$.

Definition 8 (Pareto-dominance)**.**

$$\forall a, b \in \mathcal{X}^n,\ a \succeq_P b \iff \forall i \in N\ a_i \geq b_i$$

Pareto dominance embodies the desirable property of monotonicity of a preference aggregator: if an alternative is better on every aspect than another, then it should be preferred.

Definition 9 (Lorenz vector)**.** *We call the Lorenz vector of a candidate a the cumulative vector $L(a)$ of $\mathbb{R}^n$ whose components are defined by :*

$$L(a)_i = \sum_{j=1}^{i} a_j^{\uparrow}$$

Definition 10 (Lorenz-dominance)**.** *We define the Lorenz-dominance $\succeq_L$ by :*

$$a \succeq_L b \iff \forall i \in N\ L(a)_i \geq L(b)_i$$

Example 2. *(Example 1 continued) The Lorenz vector of the three candidates $\{a, b, c\}$ are $L(a) = (0.5, 1.2, 2.2)$, $L(b) = (0.7, 1.4, 2.1)$ and $L(c) = (0.7, 1.4, 2.1)$.*
By comparing the Lorenz vectors we obtain the following Lorenz-dominance relation statements : $c \succeq_L a$ and $c \succeq_L b$.
We can note that the Lorenz-dominance is a partial preorder, as neither a or b Lorenz-dominates the other.

FOWA operators are highly linked to Lorenz Dominance as shown by Golden and Perny [6].

Proposition 1 (Reformulation from Lemma 2 in [6])**.**

$$a \succeq_L b \Leftrightarrow \forall w \in \mathcal{W}^{\searrow}\ a \succeq_w b \Leftrightarrow a \mathcal{N}_{\emptyset}^{OWA^{\searrow}} b$$

Therefore, with proposition 1 we have that the Lorenz-dominance is compatible to any FOWA operator, meaning that these results will also appear in robust FOWA but are not depending on the DM preferential information. This first set of results makes up the core of our explanatory engine described in Section 3.

3 Transitive explanations for Lorenz dominance

In this section we will present an algorithm which computes efficiently an explanation for a Lorenz dominance in the form of a transitive chain of transfers using the Pigou-Dalton Principle. The former is only a small part of the results a robust FOWA can produce and the rest will be addressed in section 4.

We saw from proposition 1 that some results, those corresponding to the Lorenz dominance $\succeq_L$, are compatible with every FOWA operator. It naturally follows that these results, will appear in the necessary preferences of any robust FOWA operator. It has also be known since the 1960s that the Lorenz dominance and the Pigou-Dalton Principle are closely related.

Definition 11 (Transitive explanation (TE))**.** *Given a set of binary relations over alternatives Y, we call transitive explanation of $a \succeq b$ using Y, a tuple $(x^0, \dots, x^{k+1}) \in (\mathcal{X}^n)^{k+2}$ such that*

$$a = x^0,\ b = x^{k+1}, \forall i \in \{0, \dots, k\} x^i\ \mathcal{R}_i\ x^{i+1},\ \textit{with } \mathcal{R}_i \in Y$$

Proposition 2 ([15], reformulation from Proposition 3.1)**.** *$a \succeq_L b$ iff there exists a transitive explanation $(x^0, \dots, x^{k+1}) \in (\mathcal{X}^{n\uparrow})^{k+2}$ using $Y = \{\succeq_{PDP}, \succeq_P\}$*

In [15], Lorenz dominance is only considered over alternatives which have the same last value in their Lorenz vectors, i.e. for which the sum of all values is the same. As we want the scope of our explanation engine to be as broad as possible, we imbue it with the capability of inserting Pareto dominance statements in the explanation sequence to overcome this limitation and deal with the potential surplus.

From proposition 2, we can build a transitive sequence of preferences, a transitive explanation, combining only progressive Pigou-Dalton transfers and Pareto dominance to explain every pair (a, b) such that $a \succeq_L b$. Note that, because of the equivalence in proposition 1, the Lorenz-dominated alternatives are exactly the ones for which explanations can solely be based on Pareto and Pigou-Dalton transfers and we will need other explanation mechanisms to explain necessary preferences of a robust FOWA operator when alternatives are not Lorenz-dominated. As Pigou-Dalton transfers only redistribute a portion ϵ among candidates without breaking the order of criteria, and as Pareto dominance is only here to remove the surplus that can remain between the last intermediate candidate and b, every intermediate candidate from $(x^0, \dots, x^{k+1})$ is within $\mathcal{X}^{n\uparrow}$.

Hence this sequence can be presented to the DM as an explanation because :

- it is of finite length;
- the mechanisms are plausible, given the explainee adheres to the principles of monotonicity and fairness they embody;

- the mechanisms used are of small cognitive load (Pareto dominance does not require any trade-off, while a Pigou-Dalton transfer can be described as occurring between two points of view, ignoring the rest); and
- the intermediate candidates used are plausible, even though they are not present in the set of candidates to rank.

The question of finding an algorithm to build a (not necessarily unique) sequence of Pigou-Dalton transfers has been answered in a close but not identical domain, on a problem called Majorization [13]. It is defined as a preorder over vectors using their values reordered in a decreasing way, so if we take similar notations as in definition 1 we would be dealing with vectors $x^{\downarrow}$. Majorization $a \succeq_{\mathcal{M}} b$ occurs when, for each criterion k, we have $\sum_{i=1}^{k} a_i^{\downarrow} \geq \sum_{i=1}^{k} b_i^{\downarrow}$. The link with our problem is therefore clear as $a \succeq_{\mathcal{M}} b \Leftrightarrow b \succeq_L a$.

In this context, the majorization is explained using a sequence of "Robin Hood transfers", which are Pigou-Dalton transfers, produced with a polynomial time algorithm. Without going into the mathematical details, we can give simply the idea of their (Lorenz-revisited) algorithm. We have $a \succeq_L b$, which means that a is more balanced than b. Especially, we can find some index j where $b_j^{\uparrow} > a_j^{\uparrow}$, and some index $k < j$ where $a_k^{\uparrow} > b_k^{\uparrow}$. The idea is to perform an exchange between these two points of view of a quantity wich is as large as possible, i.e. $\epsilon = min(a_j^{\uparrow} - b_j^{\uparrow}, b_k^{\uparrow} - a_k^{\uparrow})$. Their idea for choosing suitable values for j and k is left vague; it is usually the smallest j possible and for this j the biggest k possible.

We can pinpoint three possible drawbacks :

1. the value of ϵ does not guarantee the candidate built by the transfer to be ordered;
2. it is limited to the case where $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i$; and
3. the algorithm will find a sequence but does not aim at making it short.

Point #2 can easily be solved by allowing the explainer to use arguments based on Pareto dominance. However, this increase in flexibility makes point #3 even more prominent. Indeed, we have more flexibility in finding the criteria k so that $\sum_{i=1}^{k} a_i^{\downarrow} \geq \sum_{i=1}^{k} b_i^{\downarrow}$ as we have "surplus" (non zero Pareto-dominance implies $\sum_{i=1}^{n} a_i > \sum_{i=1}^{n} b_i$). We now present our (heuristic) Algorithm 1, which is similar to the idea given above but tries to solve the three points mentioned. A natural idea consists in minimizing the length of the explanation, but this problem seems computationally hard, even though we were not able to assess its theoretical difficulty. Consequently, we propose a heuristic method to compute short transitive explanations for a Lorenz dominance statement. In Section 6, we compare this heuristic to a A* algorithm computing an explanation of proven minimal length. Experiments tend to show we achieve nearly minimal length in a fraction of the time required to perform the exact search.

By focusing on Pigou-Dalton transfers occurring between variables which can receive (*Step 1*) or give (*Step 3 (I)*) the complete difference with the loser we ensure that the candidate obtained after the transfer is ordered. It also allows us to have a maximum length of explanation of n. Indeed at each step we remove at least one criteria from the set of criteria on which there is a non zero difference with the looser, bounding the explanation by the cardinal of this set, itself bounded by the number n of criteria.

Algorithm 1: Algorithm explaining Lorenz-dominance with Pareto dominance and Pigou-Dalton transfers

Input: $a, b \in \mathcal{X}^{n\uparrow}$ s.t. $a \succeq_L b$
Output: C
$x = a$; C $= a$
1 Compute $\mathcal{J}$, the set of indices j s.t. $x_j < b_j$ and s.t. we can perform a trade $\epsilon_j = b_j - x_j$
2 If $\mathcal{J} == \emptyset$ go to *Steps 5*
3 For each $j \in \mathcal{J}$:
(I) Compute $\mathcal{K}$, the set of indices $k < j$ s.t. $x_k > b_k$ and s.t. we can perform a trade $\epsilon_k = x_k - b_k$
(II) Find the index $k' \in \mathcal{K}$ allowing to perform the biggest trade $\epsilon = \max_{k' \in \mathcal{K}} min(\epsilon_j, \epsilon_{k'})$ (if draws take the largest index)
(III) $X = (x_1, \ldots, x_{k'} - \epsilon, \ldots, x_j + \epsilon, \ldots, x_n)$
(IV) If we don't have $X \succeq_L b$, go back to (II) to find another index in $\mathcal{K} \setminus \{k'\}$
(V) $x = X$; C $=$ C $\overset{j \to k'}{\underset{PDP}{\succeq}}$ x
4 Go back to *Step 2*
5 Compute $\mathcal{K}$, the set of indices k s.t. $x_k > b_k$
6 If $\mathcal{K} \neq \emptyset$: C $=$ C $\succeq_P b$

Once we cannot find a rank j for which the attribute value x_j is bigger than b_j, we remove every surplus that could exists with a single Pareto dominance statement. Unfortunately, our heuristic algorithm does not always leads to the smallest explanation length as shown in example 3.

Example 3. *In the same context as example 1, we want to rank students, this time over their grades in 5 courses. The two students at hand are* $d = (0.6\ 0.7\ 0.5\ 0.7\ 0.8)$ *and* $e = (0.8\ 1\ 0.6\ 0.4\ 0.4)$. *It is easy to compute that* $d \succeq_L e$, *therefore we can apply our algorithm 1 and a A* search to have two explanations.*
We reverse the explanation returned by the algorithm as it is easier to read and understand when the Pigou-Dalton transfers are performed in the reading direction. We note with $\overline{i}$ *the criterion receiving and* $\underline{i}$ *the criterion giving. We obtain for our algorithm the explanation :*

$e^{\uparrow} = (0.4\ 0.4\ 0.6\ 0.8\ 1) \preceq_P (0.4\ \overline{0.5}\ 0.6\ 0.8\ 1) \overset{5\to 2}{\underset{PDP}{\preceq}} (0.4\ \overline{0.6}\ 0.6\ 0.8\ \underline{0.9}) \overset{4\to 1}{\underset{PDP}{\preceq}} (\overline{0.5}\ 0.6\ 0.6\ \underline{0.7}\ 0.9) \overset{5\to 3}{\underset{PDP}{\preceq}} (0.5\ 0.6\ \overline{0.7}\ 0.7\ \underline{0.8}) = d^{\uparrow}$ *of length 4.*

With the A search we obtain a different explanation :*

$e^{\uparrow} = (0.4\ 0.4\ 0.6\ 0.8\ 1) \preceq_P (0.4\ 0.4\ \overline{0.7}\ 0.8\ 1) \overset{5\to 2}{\underset{PDP}{\preceq}} (0.4\ \overline{0.6}\ 0.7\ 0.8\ \underline{0.8}) \overset{4\to 1}{\underset{PDP}{\preceq}} (\overline{0.5}\ 0.6\ 0.7\ \underline{0.7}\ 0.8) = d^{\uparrow}$ *of length 3.*

In conclusion, our algorithm 1 computes in polynomial time a chain of Pigou-Dalton transfers and Pareto dominance statement to explain any Lorenz dominance statement with a bounded length of n statements. The length of the explanation is unfortunately not minimal, but the true minimal length can be computed for example with a A* algorithm over a graph exponential in size in the number of criteria.

As we saw previously, Lorenz dominance statements form a subset of the preference yielded by a robust FOWA aggregator, missing the part entailed by the specific PI obtained from the decision maker. Thus

the idea we will develop in Section 4 is to complete Pigou-Dalton transfers and Pareto dominance with a combination of statements deduced from the PI to produce a sequence of preferences, or at least a decomposition of preferences, to explain every necessary preference statements obtained by a robust FOWA operator.

4 Additive and transitive decompositions of necessary preference statements

In this section we introduce and justify the existence of decompositions for every necessary preference statement compatible with the robust FOWA $W_{\mathcal{P}}^{\searrow} \subseteq \mathcal{W}^{\searrow}$ constrained by the PI matrix $\mathcal{P}$. We start by introducing the notion of decomposition, which is weaker than the one of transitive explanation, and express the contributions of our explanation mechanisms.

The problem we want to solve is finding an explanation for the necessary preference statement $c\,\mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}}\,d$, $c, d \in \mathcal{X}^n$. We have 3 mechanisms at our disposal : PI statements provided by the DM, Pigou-Dalton transfers and Pareto dominance. Trying to find directly a valid transitive explanation $(x^0, \ldots, x^{k+1}) \in \mathcal{X}^{n\uparrow}$ using $\{\preceq_P, \preceq_{PDP}, \preceq_{\mathcal{P}}\}$ is a difficult planning problem, so we begin by building an additive decomposition of this statement.

Definition 12 (Decomposition)**.** *We call decomposition of $a \succeq b$ by Y, Y a set of explanation mechanisms, the "proto-explanation" defined by :*

$$\forall i \in N\ a_i - b_i = \sum_{y \in Y} \gamma_{yi}$$

γ_y is the contribution vector of explanation mechanism y to the preference $a \succeq b$

Remark 2. *If we take a transitive explanation $(x^0, \ldots, x^{k+1})$ of $a \succeq b$ using Y, we have $a = x^0, b = x^{k+1}$ and*
$\forall j \in \{0, \ldots, k\}\ x^j\ \mathcal{R}_j\ x^{j+1},\ \mathcal{R}_k \in Y$.
We can rewrite the latter as $x^j - x^{j+1} = \gamma_{\mathcal{R}_j}$. By summation we obtain $a - b = \sum_{j=0}^{k} x^j - x^{j+1} = \sum_{j=0}^{k} \gamma_{\mathcal{R}_j}$. Therefore an only decomposition based "proto-explanation" is weaker than a transitive explanation in the sense that any transitive explanation can always be rewritten as a decomposition.

As we have seen, invoking the anonymity of the model, we rewrite the statement $a^j \succeq_{\mathcal{P}} b^j$ by $a^{j\uparrow} - b^{j\uparrow}$ and build the $m \times n$ matrix $\mathcal{P}$ from the (transpose) m PI statements. Each line can be seen as a trade-off between criteria which is positive for the decision maker and invoking the homogeneity (and the anonymity) of the model we have $\forall x \in \mathcal{X}^n\ x^{\uparrow} + k \times (a^{j\uparrow} - b^{j\uparrow})\,\mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}}\,x^{\uparrow}$, as long as $k > 0$.

Let $a, b \in \mathcal{X}^n$ be two vectors such that $a \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} b$, determined by the GRIP method detailed in Appendix A. We show that for every such couple of candidates, we always find at least one decomposition.

Theorem 1 (PI preference decomposition)**.**

$$\begin{aligned} & a\,\mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} b \\ \Leftrightarrow\ & \exists \lambda \in \mathbb{R}_m^+ \nu, \mu \in \mathbb{R}_n^+\ s.t.\ (a^{\uparrow} - b^{\uparrow}) = \mathcal{P}^T \times \lambda + U_B^T \times \nu + \mu \\ \Leftrightarrow\ & \forall i \in N\ (a_i^{\uparrow} - b_i^{\uparrow}) = \sum_{j=1}^{m} (a_i^{j\uparrow} - b_i^{j\uparrow}) \times \lambda_j + \nu_i - \nu_{i-1} + \mu_i \end{aligned}$$

Proof. We know from the GRIP method that the minimum of $(a^{\uparrow} - b^{\uparrow})^T \times w$ is positive, entailing $a\,\mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} b$. Therefore we know that adding the constraint $(b^{\uparrow} - a^{\uparrow})^T \times w > 0$ to the sets of constraints (1), (2) and (3) will lead to an empty answer set.
From the Farkas lemma we can conclude that $-(b^{\uparrow} - a^{\uparrow})^T = (a^{\uparrow} - b^{\uparrow})^T$ can be expressed as a positive linear combination of constraints (1), (2) and (3) :

$$\begin{aligned} & \exists \lambda \in \mathbb{R}_m^+ \nu, \mu \in \mathbb{R}_n^+\ s.t. \\ & (a^{\uparrow} - b^{\uparrow})^T = \lambda^T \times \mathcal{P} + \nu^T \times U_B + \mu^T \\ & \Leftrightarrow (a^{\uparrow} - b^{\uparrow}) = \mathcal{P}^T \times \lambda + U_B^T \times \nu + \mu \end{aligned}$$

□

By an easy identifying task with the origin of the GRIP constraints, we have $\gamma_{\mathcal{P}} = \mathcal{P}^T \times \lambda$, $\gamma_{PDP} = U_B^T \times \nu$ and $\gamma_P = \mu$ as components of the decomposition for $a\,\mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} b$. We see with the sum appearing in the contribution of the PI that each PI statement contributes with a coefficient λ_j. We define a new PI-based preference relation using only one statement and its associated λ.

Definition 13 (PI dominance)**.**

$$b^{\uparrow} \overset{\lambda_j \times \mathcal{P}_j}{\underset{\mathcal{P}}{\preceq}} a^{\uparrow} \Leftrightarrow a^{\uparrow} = b^{\uparrow} + \mathcal{P}_j^T \times \lambda_j$$

The values for λ, ν, μ are most of the time not unique and can be computed with a simple linear program (n constraints corresponding to the expression of $(a_i^{\uparrow} - b_i^{\uparrow})$ in Theorem 1 and the associated domains for the variables). As our goal is to build a transitive explanation as easy as possible we need to complexify the search of values to answer the multiple objectives :

- minimize the number of PI statements involved (as it is the most cognitively demanding mechanism)
- reduce the part of Pigou-Dalton in the decomposition, *i.e.* maximize Pareto dominance
- find the nicest coefficients values required for the PI statements

To answer these objectives we decided to use a (single-)objective function $\mathcal{F}$ and to introduce a slight change in the formulation by applying an integer value $\alpha > 0$ to the left-hand side of the equation $(a^{\uparrow} - b^{\uparrow})$ and setting $lambda_j$'s to integers. In this way we can express the PI-coefficients as fractions and optimise their value. Our objective function is $\mathcal{F} = \|\lambda\|_1 + \alpha + \frac{1}{\mathcal{M}} \left\|t^-\right\|_1$ to minimize, but other approaches can be considered, such as minimising a norm L_0 or canceling balancing effects between λ and α.

5 An illustrative example

In this section we will run an example, close to a real-case study, to illustrate Algorithm 1 for Lorenz dominance and our decompositions to explain inferred preferences. Imagine taking part to a group decision with 3 others colleagues to decide the activity for the afternoon during the upcoming team-building event. As the directors board has many opportunities, the chosen activity has to be taken from a list of 7 activities : {airsoft, basketball, cycling, dancing, equestrian walk, football, golf}, which will be abbreviated with their first letters.

The idea for the team members is to provide their satisfaction rate (on a scale from 0 to 20) on the 7 activities and the board committee will chose the activity and will guaranty a sense of fairness between members. As the team is competitive, even in sports, the members

will ask for an explanation of the chosen sport, especially the most dissatisfied player who always want to know why his favorite activity was not chosen. For all these reasons the board chose (without explicitly telling the name) our robust FOWA operator.

The results of the opinion poll (reordered) for the activities to rank are :

Student	#1	#2	#3	#4
$a^{\uparrow}$	5	13	14	18
$b^{\uparrow}$	5	15	15	16
$c^{\uparrow}$	6	13	16	16
$d^{\uparrow}$	7	10	17	18
$e^{\uparrow}$	8	9	16	20
$f^{\uparrow}$	6	11	17	17
$g^{\uparrow}$	7	11	16	17

To start our model with preferences, we asked the board to rank two pairs of candidates and they replied that $b \succeq_{PI} c$ and $d \succeq_{PI} e$.

We then compute the robust FOWA operator $W_{\mathcal{P}}^{\searrow}$ and obtain the necessary preference order which is : $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, c$, $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, g$, $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, d$, $g \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, a$, $d \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, e$ and $d \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, f$, and all preferences deduced by transitivity. We can also compute the Lorenz dominance for every pair of candidates, obtaining preference statements $b \succeq_L a$, $g \succeq_L a$, $c \succeq_L a$, $c \succeq_L f$, $d \succeq_L f$.

By comparing the two preference sets, we can see that adding the preferential information $b \succeq_{PI} c$ and $d \succeq_{PI} e$ creates new preferences among candidates such as $c \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, g$ and $c \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, d$ and transitive ones such as $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, g$, $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, d$, $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, e$ and $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, f$. To sum up, our robust FOWA produces 14 preference relations, including 2 PI from the user and 12 statements to explain, 5 with Lorenz-dominance explanations and 7 with our decomposition program.

We will not detail all of these 12 statements, only one based on Lorenz-dominance such as $g \succeq_L a$ and two from our linear program such as $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, e$ and $c \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, g$. To present thing shortly we will refer as $\mathcal{P}_1 = (-1\ 2\ -1\ 0)$ and $\mathcal{P}_2 = (-1\ 1\ 1\ -1)$ the vectors corresponding respectively to the PI statements $b \succeq_{PI} c$ and $d \succeq_{PI} e$.

With Algorithm 1, we obtain as explanation for $g \succeq_L a$: $a^{\uparrow} = (5\ 13\ 14\ 18) \overset{2\to1}{\underset{P\overline{D}P}{\preceq}} (\overline{7}\ \underline{11}\ 14\ 18) \overset{4\to3}{\underset{P\overline{D}P}{\preceq}} (7\ 11\ \overline{16}\ \underline{16}) = g^{\uparrow}$.

We find two decompositions for $b \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, e$ and $c \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} \, g$:

- $b^{\uparrow} - e^{\uparrow} = (-3\ 6\ -1\ -4) = \mathcal{P}_1 + 2 \times \mathcal{P}_2 + \nu^{+} + \nu^{-}$ with $\nu^{-} = (0\ 2\ 0\ 0)$ and $\nu^{+} = (0\ 0\ -2\ 0)$
- $c^{\uparrow} - g^{\uparrow} = (-2\ 4\ -1\ -1) = \mathcal{P}_1 + \nu^{+} + \nu^{-}$ with $\nu^{-} = (0\ 1\ 0\ 0)$ and $\nu^{+} = (0\ 0\ 0\ -1)$

We were also able to find through a brute force algorithm (testing all permutation) a sequence of preferences using the decomposition to produce a transitive explanation sequence, but it is not a general result.

For $b \ \mathcal{N}_{PI}^{OWA^{\searrow}} \ e$ we found : $e^{\uparrow} = (8\ 9\ 16\ 20) \overset{3\to2}{\underset{P\overline{D}P}{\preceq}} (8\ \overline{11}\ \underline{14}\ 20) \overset{2\times\mathcal{P}_2}{\underset{\overline{\mathcal{P}}}{\preceq}} (\underline{6}\ \overline{13}\ \overline{16}\ \underline{16}) \overset{1\times\mathcal{P}_1}{\underset{\overline{\mathcal{P}}}{\preceq}} (\underline{5}\ \overline{15}\ \underline{15}\ 16) = b^{\uparrow}$

For $c \ \mathcal{N}_{PI}^{OWA^{\searrow}} \ g$ we found : $g^{\uparrow} = (7\ 11\ 16\ 17) \overset{1\times\mathcal{P}_1}{\underset{\overline{\mathcal{P}}}{\preceq}} (\underline{6}\ \overline{13}\ \underline{15}\ 17) \overset{4\to3}{\underset{P\overline{D}P}{\preceq}} (6\ 13\ \overline{16}\ \underline{16}) = c^{\uparrow}$

We can note that if we tried to apply the decomposition in another order, it would build intermediate candidates which are not valid (ordered).

6 Experimental results

In this section we will present some results obtained from experiments. We have generated randomly 1000 samples of 50 candidates (non Pareto-dominated) over 8 criteria with 5 PI statements. To compute these PI statements we draw from a Dirichlet distribution a set of ground-truth FOOWA weigth $w^{\searrow}$, and drew randomly 5 pairs (a^j, b^j) of non Lorenz-dominated candidates and added to the robust FOOWA the preference induced by $w^{\searrow}$. For the rest of the experiment the ground-truth will be hidden. From these sample we studied two aspects of our method. In the first place, we computed for Lorenz-dominated pairs of candidates the results of our algorithm 1 against A* algorithm and we compared the obtained length to the true minimal length and the computation times. In the second place, we tried through several means to compute a transitive explanation from the decompositions obtained in Section 4.

6.1 Algorithm 1 against A* algorithm

Our goal is to compare the explanation length and the computation time our polynomial heuristic algorithm 1 with a method guaranteeing to find the "true" minimum length. As introduced in Section 3 we do so by the means of a A* search. Indeed, the idea underlying the A* algorithm is to find the shortest path between the winning candidate to the loosing candidate using Pigou-Dalton transfers and Pareto-dominance. The heuristic we use to estimate the distance to the loosing candidate is the number of positions which need to receive from a Pigou-Dalton transfer, plus 1 if there are more variables in a position to give than to receive. This estimate is indeed a lower bound of the number of remaining trades.

We represent in table 1 the difference between the length found by the A* algorithm and our algorithm 1. Each row i of the table corresponds to a true length of i Pigou-Dalton transfers and Pareto-dominance. We can see that overall the results obtained by our algorithm 1 do not differ from the A* algorithm, except in some cases where up to 34% (in line 6) of the results are worst. But still, our algorithm is of interest, especially when comparing computation

Length	#Values	% identical values	Max difference
1	82 334	100 %	-
2	160 431	96.97 %	4
3	162 967	89.83 %	5
4	135 623	80.61 %	4
5	91 777	71.46 %	3
6	45 148	66.09 %	2
7	13 120	68.53 %	1
8	1539	100 %	-

Table 1. Length difference between algorithm 1 and A* algorithm.

times from table 2 and keeping in mind that during the process of finding a transitive explanation we can call for Pigou-Dalton transfers several times.

Length	#Values	Whisker	Q1	Median	Q3
1	82 334	1.37%	5.26%	6.54%	7.85%
2	160 431	11.74%	12.53%	20.10%	23.5%
3	162 967	-3.92%	20.6%	28.5%	33.61%
4	135 623	39.94%	26.54%	33.79%	39.21%
5	91 777	1.09%	38.00%	46.57%	62.05%
6	45 148	53.12%	79.18%	86.16%	91.23%
7	13 120	7.55%	93.47%	96.59%	97.96%
8	1 539	58.20%	98.10%	99.26%	99.6%

Table 2. Percentage of reduction of compute time between algorithm 1 and A* algorithm.

We have represented in table 2 the percentage of reduction in the computation time by using algorithm 1 instead of the A* algorithm. We can see that only very little data (in line 3) are corresponding to the A* algorithm performing better, and with high number of criteria we see a strict dominance of algorithm 1. Therefore, for the rest of the study combining the resolution of finding a concise explanation for Lorenz-dominance and PI dominance, we will be using algorithm 1 for its benefits in computational cost.

6.2 Feasibility of transitive explanation for decomposition computed by the MILP formulation

Our goal is to study the availability of an explanation from the decompositions defined in Section 4 for a necessary preference statement. We recall that a valid explanation is a sequence of progressive transfers from the winner to the looser, using our 3 explanation mechanisms $\{\preceq_P, \preceq_{PDP}, \preceq_{PI}\}$, while verifying that each intermediate candidate used in the explanation is valid, *i.e.* with values of criteria ordered and belonging to the domain. We will use several methods of increasing complexity to solve this problem and express the theoretical length of the explanation in terms of k the number of PI statements with a non zero associated λ_j and of l the length of Pigou-Dalton transfers to be split with Algorithm 1.

Permutation Knowing the values (λ, μ, ν), we search for a permutation verifying our validity constraints. We allow to permute each individual PI statement with the Pareto dominance and the "resulting total" Pigou-Dalton transfers. This algorithm does not try to divide any of this statements, therefore the maximum theoretical length for the explanation is $k + l + 1$. We obtain some encouraging results on the length of explanation, displayed in table 3, unfortunately too many preferences fail to have a valid permutation: we found 179 407 cases where we cannot build the transitive explanation, meaning that 74% of our explanations cannot be interpreted into valid sequences.

A first hypothesis to explain the bad results could be linked to pathological cases caused by our data sampling. Indeed, for alternatives close to the bounds of the domain $\mathcal{X}^n$, making the use of the PI statements hard without fragmentation. We therefore performed the experience again with candidates sampled over $\mathcal{X}^n = [0.1; 0.9]^n$ instead of $[0; 1]^n$, leaving this space for the explanations. The results

PI involved	#Explanations	Q1	Median	Q3
1	57 235	4	6	7
2	5 299	6	8	9
3	415	7	9	10
4	18	9	10	11
Not found	179 407	-	-	-

Table 3. Distribution of explanation length by quantity of PI involved for permutation.

are indeed better, improving the ratio to 63% of non valid sequences, but also meaning that this shape of explanation is not sufficient and should be enlarged.

Single PI Another cause could be that a part of the results are outside our validity domain, for example by the ordering constraint, due to the PI itself being highly not linear over the criteria (whereas Pigou-Dalton transfers and Pareto dominance do). Maybe some statements are required to be performed but cannot be taken separately. They could be performed in sequence but without looking into the intermediary result. Therefore our second idea is to build a unique "whole PI" statement to recompute permutation with. Even if the resulting explanation is of lesser quality, we can still advocate that its length being $l + 2$ is smaller and the harder contribution of the PI being minimized in one statement, the explanation is better for the user. Unfortunately new sequences are found but their number represent less than 1% increase in the ratio, meaning that we need to increase the complexity of our explanation pattern by splitting ν and μ to gather more explanations.

Splitting Pigou-Dalton & Pareto contributions We want to allow for more generality of the transitive pattern we want to match, in order to do so we will use the Pigou-Dalton transfer and Pareto dominance as an adjustment between uses of the PI statements. In other words we will split those transfers in order to allow for more intermediate candidates in the transitive explanation. Instead of using scheduling and planning methods directly we will step by using a intermediate MILP formulation for the chain. Indeed if we note k the number of non-zero λ_j in the decomposition, the total possible length for such a transitive explanation is $k + k * l + 1$ with $k + 1$ Pareto or Pigou-Dalton transfers. To catch some more generality, we allow for the PI statements to be used twice in the chain, therefore the total theoretical length is $2k + 2k * l + 1$ preferences in the explanation. The details of the MILP formulation are given in Appendix B.

The results in table 4 in conjunction with the results from the permutation are excellent, leaving only 16 918 explanations not found, so around 7% of the number of preferences to explain. The pattern we used here is still restrictive but with planning methods we can hope to recover almost every transitive explanation from a decomposition.

PI involved	#Explanations	Q1	Median	Q3
1	116 804	6	8	10
2	15 456	10	12	14
3	900	12	14	16
4	22	13	14.5	15.75
Not found	16 918	-	-	-

Table 4. Distribution of explanation length by quantity of PI involved for splitting transfers.

7 Conclusion and future works

In this paper we looked into the problem of producing explanations for the robust fairness-oriented OWA. The explanations are designed to be the least technical possible to be shared with external users which are possibly not involved in the conception of the system and have very little knowledge about it. To understand our explanations, it is only required to know why the model is used, preferring candidates with a balanced profile and therefore agrees with the Pigou-Dalton principle of transfers on the ordered profile of the candidate.

We presented an algorithm which builds a transitive chain for Lorenz-dominance as an explication. The existence of the chain is known from a long time [15] as a result in welfare economy and also the link with the OWA operator [6]. This algorithm works in a polynomial time in the number n of criteria, with a length of maximum n transfers. It can be used to explain Lorenz dominance in general and not only in the fairness-oriented OWA setting, where it corresponds to the results obtained by the necessary preference relation $\mathcal{N}_{\emptyset}^{OWA^{\searrow}}$.

We then presented a linear program formulation to compute an additive decomposition of a preference in $\mathcal{N}_{PI}^{OWA^{\searrow}}$, composed by a positive linear combination of the preference statements given by the DM, Pigou-Dalton transfers and Pareto-Dominance. This decomposition is not a transitive explanation, and the existence of the latter for a given decomposition is still unknown yet, even though we can always build such a decomposition (Theorem 1). However, with our experimental campaign we obtained up to 93% of available transitive explanation from the computed decomposition. We plan to adopt planning tools to reduce even further the number of non available transitive explanation (for a huge increase in computational cost) and also investigate whether guiding this search by the pre-computation of an additive decomposition is efficient or not.

REFERENCES

[1] 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). Number: 2 Publisher: Springer.

[2] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane, 'Comparing options with argument schemes powered by cancellation', in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, ed., Sarit Kraus, pp. 1537–1543. ijcai.org, (2019).

[3] Denis Bouyssou, Thierry Marchant, Marc Pirlot, Alexis Tsoukiàs, and Philippe Vincke, *Evaluation and decision models with multiple criteria: Stepping stones for the analyst*, International Series in Operations Research and Management Science, Volume 86, Boston, 1st edn., 2006.

[4] José Rui Figueira, Salvatore Greco, and Roman Slowiński, '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, (2009). Number: 2 Publisher: Elsevier.

[5] Alfio Giarlotta and Salvatore Greco, 'Necessary and possible preference structures', *Journal of Mathematical Economics*, **49**(2), 163–172, (2013). Number: 2.

[6] Boris Golden and Patrice Perny, 'Infinite order Lorenz dominance for fair multiagent optimization', in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1*, pp. 383–390, (2010).

[7] Michel Grabisch and Christophe Labreuche, 'A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid', *Annals of Operations Research*, **175**(1), 247–286, (2010). Number: 1 Publisher: Springer.

[8] Salvatore Greco, Vincent Mousseau, and Roman Slowinski, 'Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions', *Eur. J. Oper. Res.*, **191**(2), 416–436, (2008).

[9] John S. Hammond, Ralph L. Keeney, and Howard Raiffa, 'The even-swap method for multiple objective decisions', in *Research and Practice in Multiple Criteria Decision Making*, eds., Yacov Y. Haimes and Ralph E. Steuer, pp. 1–14, Berlin, Heidelberg, (2000). Springer Berlin Heidelberg.

[10] Eric Jacquet-Lagreze and Jean 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). Number: 2 Publisher: Elsevier.

[11] Ralph Keeney and Howard Raiffa. Decisions with multiple consequences: preferences and value tradeoffs, 1976.

[12] D.H. Krantz, R.D. Luce, P. Suppes, and A. Tversky, *Foundations of measurement*, volume 1: Additive and polynomial representations, Academic Press, New York, 1971.

[13] A.W. Marshall, I. Olkin, and B.C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, Springer Series in Statistics, Springer New York, 2010.

[14] Tom M Mitchell, 'Generalization as search', *Artificial intelligence*, **18**(2), 203–226, (1982). Number: 2 Publisher: Elsevier.

[15] Patrick Moyes et al., 'Gini or lorenz: Does it make a difference for inequality measurement?', in *21st Annual European Economic Association Congress-24-28 août*, (2006).

[16] Włodzimierz Ogryczak and Tomasz Śliwiński, 'On optimization of the importance weighted owa aggregation of multiple criteria', in *International Conference on Computational Science and Its Applications*, pp. 804–817. Springer, (2007).

[17] B. Roy, 'The outranking approach and the foundations of ELECTRE methods', *Theory and Decision*, **31**, 49–73, (1991).

[18] Anthony F. Shorrocks, 'Ranking Income Distributions', *Economica*, **50**(197), 3–17, (1983).

[19] Ronald R. Yager, 'On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking', in *Readings in Fuzzy Sets for Intelligent Systems*, eds., Didier Dubois, Henri Prade, and Ronald R. Yager, 80–87, Morgan Kaufmann, (1993).

A GRIP implementation

The GRIP method described by Figueira et al. [4] allows to represent any additive utility function of a vector x by $U(x) = \sum_{i=1}^{n} u_i(x_i)$ with u_i the non decreasing marginal utility function of criterion i. The framework allows the marginals to be piece-wise linear.

From a learning set $A^R \subseteq \mathcal{X}^n \times \mathcal{X}^n$ corresponding to inequality constraints between the utilities of vectors, a set of inequalities to guaranty the piece-wise linear shape and non decreasingness of the marginals and a constraint on the domain of U (*i.e.* sum of weights equals to one), they define a set of constraints $(E(x, y))$ for any pair (x,y) of candidates. If when optimising the function $d(x, y) = min\{U(x) - U(y)\}$ we obtain a positive value, then we can conclude that $x \succeq_N y$, x is necessarily preferred over y (section 5 of [4].

The GRIP method is very general and allows to represent complex shape for the utility function but in our OWA case some of them are unnecessary :

- our marginal utility functions are linear, therefore the second set of constraints described collapses into the non negativity of the OWA weights w_i
- in Section 6 of [4] they propose to implement intensity of preferences, thing we won't use in this paper

We will also add new constraints on the decreasingness of w as explained in Section 2. The resulting linear program is composed by sets of contraints (1)-(3) :

$$\left.\begin{array}{l}\sum_{i=0}^{n} w_i = 1 \\ w \geq 0\end{array}\right\} \Leftrightarrow w \in \mathcal{W} \tag{1}$$

$$U_B \times w \geq 0 \Leftrightarrow w \in \mathcal{W}^{\searrow} \tag{2}$$

$$\begin{array}{l}\forall j \in \{1, \ldots, m\}\ (a^{j\uparrow})^T \times w \geq (b^{j\uparrow})^T \times w \\ \qquad \Leftrightarrow \mathcal{P} \times w \geq 0 \Leftrightarrow w \in W_{PI}^{\searrow}\end{array} \tag{3}$$

with U_B the upper diagonal matrix equal to $\begin{pmatrix} 1 & -1 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \ldots & 0 & 1 \end{pmatrix}$ and $\mathcal{P}$ the $m \times n$ matrix containing $(a^{j\uparrow} - b^{j\uparrow})^T$.

The objective function $d(x, y)$ is simply

$$d(x, y) = min\{(x - y)^T \times w\}$$

and if we obtain a positive result, we can deduce $x \, \mathcal{N}_{\mathcal{P}}^{OWA^{\searrow}} y$

B MILP formulation for transitive explanation reconstruction

The idea of this third method to reconstruct the transitive explanation is to alternate Pigou-Dalton & Pareto with the use of a PI statement. As we also know the number k of PI statement, we know the total number of intermediate candidate we have to put inside our linear program : $4k + 1$, corresponding to the alternating $2k$ PI statements with $2k$ Lorenz dominance (regrouping Pigou-Dalton and Pareto dominance as presented in Section 3. We decided not to impose the use of either Pigou-Dalton transfer or Pareto dominance at each step to reduce the use of binaries and therefore the complexity of the MILP to solve.

To keep the notations from the transitive explanation, we will have intermediate candidates $(x^0, \ldots, x^{4k})$, with $x^0 = a$ and $x^{4k} = b$. From the alternating sequence we have :

1. $\forall j \in \{0, \ldots, 4k - 2\},\ j = 2m,\ x^j \succeq_L x^{j+1}$
2. $\forall j \in \{1, \ldots, 4k - 1\},\ j = 2m + 1,\ x^j \underset{\mathcal{P}}{\overset{\lambda_i \times \mathcal{P}_i}{\succeq}} x^{j+1}$
3. $x^{4k-1} \succeq_L x^{4k}$

Before giving the constraints for this set of constraint, we will first ensure with equation (4) that each intermediate candidate is ordered and belongs to $\mathcal{X}^n$.

$$\forall j \in \{0, \ldots 4k\}\ x^j \in \mathcal{X}^n,\ U_B^T \times x^j \geq 0 \tag{4}$$

Item (1) and (2) are composed by a part of the right-hand in the equation in Theorem 1, and give the sets of equations (5) and (6)-(9).

$$\begin{array}{l}\forall j \in \{0, \ldots, 4k - 2\},\ j = 2m,\ (x^j - x^{j+1}) = U_B^T \times \nu^m + \mu^m \\ \qquad \nu, \mu \in M_{n,2k}(\mathbb{R}^+)\end{array} \tag{5}$$

The equations for the PI statement formatting are harder and require the use of binary variables.

$$\begin{array}{l}\forall j \in \{1, \ldots, 2k - 1\},\ j = 2m + 1,\ \forall i \in N \\ \qquad (x_i^j - x_i^{j+1}) = \sum_{j \in M} \mathcal{P}_i^j \times \lambda_j \times \gamma_j^m\end{array} \tag{6}$$

$$\forall j \in \{1, \ldots, k\} \Gamma^j - \gamma^j \geq 0 \tag{7}$$

$$\forall j \in \{1, \ldots, k\} \left\|\Gamma^j\right\|_1 \leq 1 \tag{8}$$

$$\forall i \in N,\ \sum_{j=1}^{k} \Gamma_i^j \leq 1 \tag{9}$$

$$\gamma \in M_{m,k}([0, 1]),\ \Gamma \in M_{m,k}(\{0, 1\})$$

In equation (6), we recall that λ is given and is no more a variable in this problem, so the constraint is linear. The γ_j^m variable is the fragment of the PI statement j that will be used in the preference $x^{2m+1} \underset{\mathcal{P}}{\overset{\gamma_j^m \lambda_j \times \mathcal{P}_j}{\succeq}} x^{2m+2}$. By using a binary Γ^m we ensure with equation (8) that at most one PI statement is used for this preference m and with equation (9) one statement j is used only once in the first

half of possible PI statements.

We need to write the same equations (6)-(9) with $j \in \{2k+1, \ldots, 4k-1\}$ to constrain the PI statements for the second half of the possible PI statements and add equation (10) to use each λ_j completely.

$$\forall i \in N, \sum_{j=1}^{2k} \gamma_i^j = 1 \tag{10}$$

To complete our mixed integer linear programming formulation we need to discuss the objective function. Our goal is to limit at most the split of the PI statement, as their number and their use is fixed. Limiting the use of Pigou-Dalton statements could not be performed with a MILP formulation as the true length hidden behind the Lorenz dominance is given by Algorithm 1. To conclude the function we want to minimize is :

$$\sum_{j=1}^{2k} \left\| \Gamma^j \right\|_1 + \sum_{j=1}^{2k} \left\| \nu^j \right\|_1$$

even if the second part should be improved and coefficients introduced to break the symmetry of answers and help the overall convergence.