Principled Explanations for Robust Redistributive Decisions

Hénoïk Willot, Khaled Belahcene, Sébastien Destercke

Principled Explanations for Robust Redistributive Decisions

Hénoïk Willot[a,*], **Khaled Belahcene**[b] **and Sébastien Destercke**[a]

[a]Heudiasyc, University of Technology of Compiègne, France
[b]MICS, CentraleSupélec, Université Paris-Saclay, France

Abstract. Ordered weighted averaging (OWA) functions, a.k.a. Generalised Gini Index, are routinely used to obtain fair solutions. However, while they ensure some level of fairness in the result, two remaining questions are why they recommend a given alternative, and how robust is this recommendation. We bring practical and theoretically grounded solutions to these questions, by providing an explanation engine for robust OWA that consists in a normative transitive chain of self-evident arguments, themselves based on the normative properties of the model. We provide a thorough theoretical study of the engine, showing that it is sound and complete with respect to the model, with a theoretical upper bound on the length of the explanation and a tractable algorithm (even though minimizing the length is NP-hard). We also provide experimental evidence that the engine performs well on synthetic data. Thus, we guarantee that an explanation can always be found, and that reasoning according to the provided scheme always produces a valid statement. Moreover, the explanations allow to probe the normative requirements of the model, so as to allow validation, accountability and recourse, that are key components of trustworthy AI.

1 Introduction

Fairness, Robustness and Explainability are three pillars of trustworthy AI. *Ordered weighted averages* (OWA) [43], a.k.a. *generalized Gini indices* [41] are commonly used to ensure the first requirement of fairness in different settings such as economy and computational social choice [2], multi-objective optimisation [8], or preference learning [18] to name a few. However, we are unaware of works that consider explaining robust fair OWA, that is OWA operators with decreasing weights defined by partial information. This contrasts with other popular models of comparable complexity such as robust weighted averages, for which sound, complete and efficient explanation engines exist.

We give answers to this issue, first by characterising robust OWAs through normative properties, and then by using this characterisation to produce demonstrably sound, complete, and algorithmically efficient explanation engines. We have several reasons to adopt such a normative, logical viewpoint:

- it paves the way to certified and accountable systems, allowing parties feeling prejudiced to dispute a recommendation result on formal grounds;
- OWAs are not amenable to approximate/statistical explanation tools such as Shapley values [29] (as they are symmetric) or gradient information [36] (as they are highly non-linear);
- produced explanations rely on the provided information and normative properties, without extra assumptions or a disclosure of the estimated parameters. This makes breach of privacy or manipulation more difficult to perform, and minimizes the inductive bias.

Concretely, we strive to explain *comparative statements* of the type "$\mathbf{x}$ is less desirable than $\mathbf{y}$", where $\mathbf{x}, \mathbf{y}$ are alternatives or states of the world, and where our decision function has to satisfy a number of desirable normative properties together with observed preferences. Furthermore, we base our conclusions and explanations on deductions valid for every possible model consistent with our information. Such skeptic inferences are routinely used in logic [19] as well as in multi-objective [1] and uncertain [38, 30] decision problems, ensuring robustness in a strong sense.

We will build up our proposal by starting from basic normative properties and climb the ladder of complexity towards robust OWA (which, as a limit, include precise ones). In the process, our results will point to a reason of why explaining fair robust OWAs has received little to no attention: the problem is not trivial, both from a theoretical and algorithmic perspective, and is not merely an adaptation of results existing for, say, the weighted average. As a starting point, we assume alternatives are described along a number of viewpoints on commensurate scales, and that comparative statements correspond to a preference structure satisfying strong decision-theoretic properties (transitivity, symmetry, redistributivity and monotonicity) that are normatively desirable for the decision process under scrutiny. Such preference structures are at work in the notion of generalized Lorenz dominance, introduced a long time ago by welfare economists [39].

In section 4, we proceed to fair robust OWA (i.e., with decreasing weights), that refines such generalized Lorenz-dominance, solving the issue that this preference relation can often result in incomparabilities between alternatives, i.e. is too indecisive. In particular, we consider that in addition to satisfying the previously mentioned decision-theoretic properties, a user has provided preference statements, for instance through some active learning scheme [7], but that those statements only allow to identify a subset of possible models.

Our main contributions are the following :

- We improve upon the existing literature about explanations of Lorenz-dominance: we show that finding optimal (i.e., shortest) explanations consisting of sets of successive transfers is an NP-

* Corresponding Author. Email: henoik.willot@hds.utc.fr

hard problem, and provide heuristics that are empirically better than previous ones.

- We provide what is to our knowledge new axiomatics for convex sets of OWA (which we refer to as robust fair OWA), that embeds Lorenz dominance and generalised Gini index in a single, unifying framework. Moreover, this axiomatic being rather natural, it allows us to provide explanations mechanisms. It contrast with axiomatics relying on technical axioms such as continuity, that are difficult to leverage into explanations. We also show that our explanations are logically sound and complete, meaning that all preferences can be explained, and all explained preferences are true. We also provide heuristics to provide such explanations rapidly.

Further details such as proofs, optimisation problems and their implementation, and experimental results can be found in Appendix.

2 Basic Notions

We consider preferences between *alternatives* described along several attributes measured on a common continuous scale: we denote by $[n] = \{1, \dots, n\}$ the set of attributes, and by $\mathbb{X}$ this common scale, $\mathbb{X}$ being a non-trivial interval of the real line. Alternatives $\mathbf{x} \in \mathbb{X}^n$ are denoted by small case letters. $[n]$ can represent viewpoints in different settings, such as multiple criteria decision-making or multi-agent frameworks, and alternatives are options described along those viewpoints, for instance distribution of wealth among agents. We denote $(\mathbf{e}^1, \dots, \mathbf{e}^n)$ the canonical base of $\mathbb{R}^n$ and $\widehat{\mathbb{X}^n}$ the subset of $\mathbb{X}^n$ of tuples sorted in non-decreasing order.

Preference is represented by a binary relation $\mathcal{R}$ on $\mathbb{X}^n$. Given two alternatives $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{X}^n$, the *comparative statement* $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ (or $\mathbf{x}\ \mathcal{R}\ \mathbf{y}$) denotes that alternative $\mathbf{x}$ is at most as desirable as alternative $\mathbf{y}$. Consequently, there are four possible outcomes when comparing $\mathbf{x}$ to $\mathbf{y}$: *(i)* $\mathbf{y}$ is strictly preferred to $\mathbf{x}$ if $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ and $(\mathbf{y}, \mathbf{x}) \notin \mathcal{R}$; *(ii)* $\mathbf{x}$ is strictly preferred to $\mathbf{y}$ if $(\mathbf{x}, \mathbf{y}) \notin \mathcal{R}$ and $(\mathbf{y}, \mathbf{x}) \in \mathcal{R}$; *(iii)* $\mathbf{x}, \mathbf{y}$ are indifferent when $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ and $(\mathbf{y}, \mathbf{x}) \in \mathcal{R}$[1]; *(iv)* $\mathbf{x}, \mathbf{y}$ are *incomparable* when $(\mathbf{x}, \mathbf{y}), (\mathbf{y}, \mathbf{x}) \notin \mathcal{R}$.

Some preference statements are of specific interest:

Definition 1 (reorderings $\mathcal{S}$)**.** Let $\mathcal{S}$ be the set of comparative statements $(\mathbf{x}, \mathbf{y})$ s.t. $\mathbf{y}$ is a permutation of $\mathbf{x}$. Obviously, $\mathcal{S}$ is an equivalence relation, and every alternative $\mathbf{x} \in \mathbb{X}^n$ has a unique equivalent $\widehat{\mathbf{x}}$ w.r.t. $\mathcal{S}$ in the set $\widehat{\mathbb{X}^n}$.

Definition 2 (transfers $\mathcal{T}$)**.** With $t \in \mathbb{R}$ and $i, j \in [n]$, let $\tau^t_{j \to i}$ the comparative statement $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$ where both alternatives are sorted in non-decreasing order, and $\widehat{\mathbf{y}}$ is the situation where, starting with $\widehat{\mathbf{x}}$, the amount t is taken from agent j and given to an agent i, i.e. $\widehat{\mathbf{y}} = \widehat{\mathbf{x}} + t\mathbf{e}^i - t\mathbf{e}^j$. When $t > 0$ and $i < j$, this transfer is called *redistributive*[2], and we denote $\mathcal{T}$ the set of all redistributive transfers, i.e. $\mathcal{T} := \bigcup_{t>0} \bigcup_{1 \le i < j \le n} \tau^t_{j \to i}$.

Definition 3 (gifts $\mathcal{G}$)**.** Let $\mathcal{G} := \{(\mathbf{x}, \mathbf{y}) : \forall i \in [n]\ \mathbf{x}_i \le \mathbf{y}_i\}$, the set of comparative statements where the preference of the agents in $[n]$ is unanimous[3]

We give a global example where a central authority has to dispatch investment revenues, illustrating both the considered problem, the involved notions and the provided solutions. We invite readers to come back to it along their reading.

[1] Obviously, this requires that $\mathcal{R}$ is *reflexive*, i.e. $(\mathbf{x}, \mathbf{x}) \in \mathcal{R}$. We silently make this assumption throughout the paper.

[2] One can see that as "take an amount $t > 0$ from the richer agent j and give it to the poorer one i" (while maintaining the social order). Such transfers are known as *Pigou-Dalton* or *Robin Hood* transfers.

[3] This is simply Pareto-dominance of $\mathbf{x}$ by $\mathbf{y}$

Example 1 (illustrative example)**.**

Investment	*Alice*	*Bob*	*Charlie*	*David*	*Emma*
$\mathbf{a}$	*31*	*83*	*70*	*16*	*51*
$\mathbf{b}$	*28*	*98*	*25*	*2*	*84*
$\mathbf{c}$	*22*	*23*	*76*	*82*	*34*
$\mathbf{d}$	*96*	*6*	*18*	*17*	*88*

We begin by claiming alternatives should first be anonymized and sorted by increasing order of satisfaction.

Investment	*#1*	*#2*	*#3*	*#4*	*#5*
$\widehat{\mathbf{a}}$	*16*	*31*	*51*	*70*	*83*
$\widehat{\mathbf{b}}$	*2*	*25*	*28*	*84*	*98*
$\widehat{\mathbf{c}}$	*22*	*23*	*34*	*76*	*82*
$\widehat{\mathbf{d}}$	*6*	*17*	*18*	*88*	*96*

Suppose the central authority has already stated that $\mathbf{b}$ *is at most as desirable as* $\mathbf{d}$ *and is using a precise OWA operator*[4] $\mathcal{O}_{\mathrm{WA}_{\{\boldsymbol{\omega}^*\}}}$ *defined by the value of its parameter* $\boldsymbol{\omega}^* = (.70, .10, .10, .05, .05)$*. How can they explain that* $\mathbf{c}$ *is preferred to* $\mathbf{a}$ *while keeping* $\boldsymbol{\omega}^*$ *hidden?*[5]

To prove this statement, we build a chain of alternatives $(\mathbf{a}, \widehat{\mathbf{a}}, \mathbf{x}^1, \mathbf{x}^2, \widehat{\mathbf{c}}, \mathbf{c})$*, with* $\mathbf{x}^1 := (16\ 35\ 47\ 70\ 83)$ *and* $\mathbf{x}^2 := (22\ 23\ 32\ 76\ 80)$*.* $\mathbf{x}^1$ *should be considered as better than* $\widehat{\mathbf{a}}$*, because it is obtained by transferring* 4 *units from the third least satisfied agent to the second least satisfied. Situation* $\mathbf{x}^2$ *should be considered as better than* $\mathbf{x}^1$*, as the change* $(+6, -12, -15, +6, -3)$ *from* $\mathbf{x}^1$ *to* $\mathbf{x}^2$ *should be considered positive, as it corresponds* ceteris paribus *to one and a half time the change from* $\widehat{\mathbf{b}}$ *to* $\widehat{\mathbf{d}}$*, considered to be positive by the central authority. Finally,* $\widehat{\mathbf{c}}$ *should be considered as better than* $\mathbf{x}^2$*, because every agent is at least as satisfied. Thus, transitivity of preference leads to preferring* $\mathbf{c}$ *over* $\mathbf{a}$*.*

Technically, the statements $(\mathbf{a}, \widehat{\mathbf{a}})$ *and* $(\widehat{\mathbf{c}}, \mathbf{c})$ *are reorderings,* $(\widehat{\mathbf{a}}, \mathbf{x}^1)$ *is a redistributive transfer and* $(\mathbf{x}^2, \widehat{\mathbf{c}})$ *is a gift.*

3 Explaining Lorenz-Dominance Statements

Our goal will be to characterise skeptic inferences and explanation mechanisms for fair decision rules, starting with the restricted Lorenz-Dominance and then proceeding to the generalized version, whose definitions are recalled below

Definition 4 (Lorenz dominance relations $\mathcal{L}$ and $\mathcal{L}^*$)**.** The *generalized Lorenz dominance* is the binary relation $\mathcal{L}$ over $\mathbb{X}^n$ such that $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}$ iff $\forall i \in [n], \sum_{k=1}^{i} \widehat{\mathbf{x}}_k \le \sum_{k=1}^{i} \widehat{\mathbf{y}}_k$. The *restricted Lorenz dominance* is the subset $\mathcal{L}^*$ of $\mathcal{L}$ where $\mathbf{x}, \mathbf{y}$ have the same total income, i.e. $\sum_{k \in [n]} \mathbf{x}_k = \sum_{k \in [n]} \mathbf{y}_k$.

3.1 The Semantics of Skeptical Preference

In this paper, we are interested in preference relations satisfying a number of properties coming from Decision Theory.

Property (t). $\mathcal{R}$ is *transitive* when, for all alternatives $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{X}^n$, if $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ and $(\mathbf{y}, \mathbf{z}) \in \mathcal{R}$, then $(\mathbf{x}, \mathbf{z}) \in \mathcal{R}$.

Property (s). $\mathcal{R}$ is *symmetric* when it extends (or refines) $\mathcal{S}$, i.e. $\mathcal{R} \supseteq \mathcal{S}$ contains all reorderings.

[4] The numerical representation is recalled in preamble of Section 4.1.

[5] We note that computing individual agents importance e.g. of Bob with the partial derivative $\partial_{\mathrm{Bob}} \mathcal{O}_{\mathrm{WA}_{\{\boldsymbol{\omega}^*\}}}(\mathbf{a}) = .70$, attributes high importance to agent Bob and is thus misleading, while the Shapley values of players Alice, Bob, Charlie, David and Emma are equal and thus uninformative.

Property (r). $\mathcal{R}$ is *redistributive* when it extends $\mathcal{T}$, i.e. $\mathcal{R} \supseteq \mathcal{T}$ contains all redistributive transfers.

We denote $\mathfrak{P}_{(t,s,r)}$ the set of all reflexive binary relations over $\mathbb{X}^n$ satisfying simultaneously these three properties. Assuming this set is not empty (this is proven in next section), let $\mathcal{P}_{(t,s,r)}$ denote the preference relation defined as the intersection of all relations in $\mathfrak{P}_{(t,s,r)}$. As the three properties are stable under intersection[6], $\mathcal{P}_{(t,s,r)}$ belongs to $\mathfrak{P}_{(t,s,r)}$ and is its smallest element w.r.t. set inclusion. Thus, semantically, $\mathfrak{P}_{(t,s,r)}$ can be seen as the set of every possible world, and $\mathcal{P}_{(t,s,r)}$ as the set of comparative statements that can be skeptically inferred, i.e. that hold in every possible world. $\mathfrak{P}_{(t,s,r)}$ can be seen as a jury, where the jurors are all the preference relation satisfying the normative properties, and $\mathcal{P}_{(t,s,r)}$ is the set of unanimous decisions among them. Those cautious decisions are necessary w.r.t. the normative principles, and are not subject to arbitrariness or contingency. By focusing on these decisions, we offer robustness to the user, and we hope to support them with proofs and explanations grounded on the normative properties.

3.2 A Numeric Representation of $\mathcal{L}^*$

It is easy to check that both $\mathcal{L}$ and $\mathcal{L}^*$ given in Def. 4 satisfy properties (t), (s) and (r). Hence, the set $\mathfrak{P}_{(t,s,r)}$is not empty, and in fact we shall see that $\mathcal{L}^*$ is its smallest element, i.e., $\mathcal{P}_{(t,s,r)} = \mathcal{L}^*$. This is a powerful result, meaning that checking if a given comparative statement $(\mathbf{x}, \mathbf{y})$ holds for every preference relation that satisfies (t), (s) and (r), i.e. is a *robust* decision under these assumptions, one simply needs to sort $\mathbf{x}$ and $\mathbf{y}$, compute their cumulative sums and perform (at most) n element-wise comparisons.

3.3 A Sound and Complete Calculus of Preference

We are interested in building a formal deductive system that mirrors the decision-theoretic properties, allowing to infer comparative statements from tuples of comparative statements.

Inference rule. We associate property (t) to the rule

$$\textbf{Rule T :}\ \frac{(\mathbf{a},\mathbf{b}),(\mathbf{b},\mathbf{c})}{(\mathbf{a},\mathbf{c})}\ \text{(transitivity)}$$

Basic truths[7]**.** To reflect the properties (s) and (r), we consider reorderings $\mathcal{S}$ and redistributive transfers $\mathcal{T}$ as self-evident statements.

Let $cl_T(\mathcal{S} \cup \mathcal{T})$ be the deductive closure[8] of the set of basic truths $\mathcal{S} \cup \mathcal{T}$ under the operator T, i.e. the set of all comparative statements that can be proved from premises in $\mathcal{S}$ or in $\mathcal{T}$ by chaining deductions according to the rule T. *Soundness* of the formal system w.r.t the semantics, i.e. $cl_T(\mathcal{S} \cup \mathcal{T}) \subseteq \mathcal{P}_{(t,s,r)}$, meaning that every proven statement skeptically holds, immediately follows from the construction of the rules and axioms mirroring the properties of preference. *Completeness*, i.e. $\mathcal{P}_{(t,s,r)} \subseteq cl_T(\mathcal{S} \cup \mathcal{T})$, meaning that every statement that cannot be empirically disproved can indeed be proved, follows from the fact $cl_T(\mathcal{S} \cup \mathcal{T})$ satisfies (t) because it is closed under T, and obviously (s) and (r).

[6] in the sense that if $\mathcal{R}_1$ and $\mathcal{R}_2$ both satisfy this property, then so does $\mathcal{R}_1 \cap \mathcal{R}_2$.

[7] A.k.a *axioms*, but we eschew terms whose meaning changes whether they are used in Logics or Decision theory.

[8] As the sole operator here is transitivity, it is also the transitive closure, but this is contingent to the properties of restricted Lorenz dominance.

3.4 Schematic Explanations

As satisfying as the completeness result is, proofs resulting from are still in the form of trees, which are likely to not be concise nor simple enough to be cognitively accepted by agents. Hence we propose explanations shaped as transitive *argument schemes* forming a sequence of "speech acts". This formalism allows the representation of varied types of explainees and explanation situations: while experts can be given trees, lay persons might require simpler arguments.

Definition 5 (ATX scheme)**.** Let $\mathcal{R}$ be a binary relation over $\widehat{\mathbb{X}^n}$. An *anonymous-transitive explanation based on evidence in* $\mathcal{R}$ of length ℓ ($\mathcal{R}$-ATX$^\ell$) is a pair (s, c) where the *support* s is a ℓ-tuple of comparative statements $s = ((\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^\ell, \mathbf{y}^\ell)) \in \mathcal{R}^\ell$ and the *claim* c is a comparative statement $c = (\mathbf{x}, \mathbf{y})$ satisfying $\mathbf{x}^1 = \widehat{\mathbf{x}}$, $\mathbf{y}^\ell = \widehat{\mathbf{y}}$ and, for all integers k between 2 and ℓ, $\mathbf{x}^k = \mathbf{y}^{k-1}$.

Let $\ell \in \mathbb{N}^*$ and $\mathcal{E}(\mathcal{T}\text{-ATX}^\ell)$ denote the set of all *explainable statements*, i.e. claims associated to a chain of self-evident statements in $\mathcal{T}$ by some ATX of length ℓ. This set is included in $cl_T(\mathcal{S} \cup \mathcal{T})$, because an ATX supporting the conclusion $(\mathbf{x}, \mathbf{y})$ can be seen as a transitive chain between $\mathbf{x}$ and $\mathbf{y}$, where the first (i.e. $(\mathbf{x}, \widehat{\mathbf{x}})$) and last (i.e. $(\widehat{\mathbf{y}}, \mathbf{y})$) comparative statements are in $\mathcal{S}$, and every other statement is in $\mathcal{T}$: $\mathcal{T}$-ATX are sound w.r.t. $cl_T(\mathcal{S} \cup \mathcal{T})$. Are they complete? Indeed, they are even complete w.r.t. $\mathcal{L}^*$, from a well-known result from 1929.

Lemma 1 (Hardy, Littlewood and Polya [23])**.** $\mathcal{L}^* \subseteq \bigcup_{\ell=1}^{n} \mathcal{E}(\mathcal{T}\text{-}ATX^\ell)$

We briefly recall the constructive proof, as it forms an algorithm we intend to modify and build upon.

Sketch of proof. Initialize $\mathbf{x}^0 := \widehat{\mathbf{x}}$. At step k find i^*, j^*, t^* s.t. $j^* = \arg\min_j \mathbf{x}_j^{k-1} > \widehat{\mathbf{y}}_j$, $i^* = \arg\max_{i:i<j} \mathbf{x}_i^{k-1} < \widehat{\mathbf{y}}_i$ and $t^* = \min(\widehat{\mathbf{y}}_{i^*} - \mathbf{x}_{i^*}^{k-1}, \mathbf{x}_{j^*}^{k-1} - \widehat{\mathbf{y}}_{j^*})$ then define $\mathbf{x}^k$ s.t. $(\mathbf{x}^{k-1}, \mathbf{x}^k) \in \tau_{j^* \to i^*}^{t^*}$. The number of agents $i \in [n]$ s.t. $\mathbf{x}_i^k \neq \widehat{\mathbf{y}}_i$ strictly decreases with k, hence the algorithm terminates and provides a $\mathcal{T}$-ATX of length at most n for $(\mathbf{x}, \mathbf{y})$. □

To summarise, the various soundness results together with the completeness of explanations based on redistributive transfers w.r.t. restricted Lorenz dominance amount to

Theorem 1. $\bigcup_{\ell=1}^{n} \mathcal{E}(\mathcal{T}\text{-}ATX^\ell) = cl_T(\mathcal{S} \cup \mathcal{T}) = \mathcal{P}_{(t,s,r)} = \mathcal{L}^*$

Example 2. *Consider alternatives* $\widehat{\mathbf{c}}$ *and* $\widehat{\mathbf{b}}$ *of Example 1. Computing their cumulative sums show that* $(\mathbf{b}, \mathbf{c}) \in \mathcal{L}^*$*. The shortest explanation supporting* $(\mathbf{b}, \mathbf{c})$ *has length 4:*

$$\widehat{\mathbf{b}}\ \tau_{4\to1}^{18}\ (\overline{20}\ 25\ 28\ \underline{66}\ 98)\ \tau_{5\to4}^{10}\ (20\ 25\ 28\ \overline{76}\ \underline{88})$$

$$\tau_{2\to1}^{2}\ (\overline{22}\ \underline{23}\ 28\ 76\ 88)\ \tau_{5\to3}^{6}\ \widehat{\mathbf{c}}$$

3.5 Computational Aspects

According to Th. 1, given a comparative statement $(\mathbf{x}, \mathbf{y})$, it is equivalent to *(i)* decide if it holds for every preference structure satisfying the properties (t), (s) and (r); *(ii)* search for a deductive proof following the rule T with premises in $\mathcal{S}$ and $\mathcal{T}$; *(iii)* solve the problem of finding an explanation; or *(iv)* sort out - sum up - compare in accordance to the numeric representation of $\mathcal{L}^*$. Obviously, the latest is the easiest from the computational viewpoint. Indeed, we prove the problem of finding an explanation of a given size is intractable.

$$\mathcal{D}_k^* = \left\{ j \in [n] \mid \mathbf{x}_j^{k-1} \geq \widehat{\mathbf{y}}_j \text{ and } \mathbf{x}_j^{k-1} - \mathbf{x}_{j-1}^{k-1} \geq \mathbf{x}_j^{k-1} - \widehat{\mathbf{y}}_j \right\},$$
$$\mathcal{R}_k^* = \left\{ i \in [n] \mid \mathbf{x}_i^{k-1} \leq \widehat{\mathbf{y}}_i \text{ and } \mathbf{x}_{i+1}^{k-1} - \mathbf{x}_i^{k-1} \geq \widehat{\mathbf{y}}_i - \mathbf{x}_i^{k-1} \right\},$$
$$t_k^* = t(i_k^*, j_k^*) = \max_{i \in \mathcal{R}_k^*, j \in \mathcal{D}_k^*, i<j} \min \left(\mathbf{x}_j^k - \widehat{\mathbf{y}}_j, \widehat{\mathbf{y}}_i - \mathbf{x}_i^k \right)$$

Figure 1. Contribution algorithm– choice of donor j^*, receiver i^* and amount t^* for the transfer at step k.

Theorem 2 (Hardness of finding short explanations)**.** *Given $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}^*$ and a positive integer k, deciding if there is a $\mathcal{T}$-ATX of length at most k for $(\mathbf{x}, \mathbf{y})$ is NP-hard.*

Sketch of proof. We propose a reduction from the 3-partition problem [20]. Let the set S of integers such that $|S| = 3m$, $\sum_{s \in S} s = mT$ and $\frac{T}{4} < s < \frac{T}{2}$ $\forall s \in S$ be the input of the 3-partition problem. We build the alternatives $\mathbf{x}$ and $\mathbf{y} \in \widehat{\mathbb{X}^{4m}}$ such that $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}^*$, defined by $\mathbf{x}_i = \sum_{j=1}^{i-1} \widehat{S_j}$ if $1 \leq i \leq 3m$ and $\mathbf{x}_i = iT$ if $3m < i \leq 4m$ and by $\mathbf{y}_i = \sum_{j=1}^{i} \widehat{S_j}$ if $1 \leq i \leq 3m$ and $\mathbf{y}_i = (i-1)T$ if $3m < i \leq 4m$. □

This intractability result leads us to consider:

- A mixed integer linear optimization (MILO) formulation in terms of a continuous planning problem[9] whose solution is the shortest possible explanation.
- An efficient greedy heuristic, akin to the folk algorithm by Hardy, Littlewood and Polya (HLP) [23] underlying Lemma 1, but geared towards brevity is described in Figure 1 by its choice of values for i^*, j^*, t^* for step k.

Experimentation results on synthetic data are given in Table 1. For a given number of agents n, a population of 10 alternatives is sampled from $\mathbb{X}^n := [1000]^n$ s.t. the total income of each alternative is equal to $200n$ (more details are given in Section A.4 and the corresponding code is available [42]). The results presented, obtained on an ordinary laptop, are averaged over 10 independent repetitions. Table 1 shows that our heuristic is very fast and close to the optimal answer. Results suggest optimal length grows as $0.6n$, while our heuristic have a $0.7n$ length growth.

3.6 The Case of Generalized Lorenz-Dominance

The machinery we have just built does not allow to compare populations having a different total outcome. For this reason, economists have proposed to consider *gifts* as desirable.

Property (m). $\mathcal{R}$ is *monotonic* when $\mathcal{R} \supseteq \mathcal{G}$.

This property can be seen as a guarantee of *efficiency*: spending the surplus is preferred to not spending it[10].

Semantics and Representation. We observe that $\mathcal{L}$ satisfies (m) (while $\mathcal{L}^*$ does not), hence, the set $\mathfrak{P}_{(t,s,r,m)}$ of all reflexive binary relations satisfying conjointly (t), (s), (r) and (m) is not empty. Let $\mathcal{P}_{(t,s,r,m)}$ be the intersection of all relations in $\mathfrak{P}_{(t,s,r,m)}$, which is the smallest element of $\mathfrak{P}_{(t,s,r,m)}$, as monotonicity is also stable under intersection. $\mathfrak{P}_{(t,s,r,m)}$is a (strict) subset of $\mathfrak{P}_{(t,s,r)}$, thus $\mathcal{P}_{(t,s,r,m)}$ refines $\mathcal{P}_{(t,s,r)}$: considering an additional property narrows the spectrum of possible worlds and reduces the availability of counter-arguments to a preference claim, thus allowing more comparative statements to skeptically hold.

Deduction. Accounting for (m) in the deductive system is simple: it suffices to consider the gifts $\mathcal{G}$ as basic truths, in addition to the reorderings $\mathcal{S}$ and the transfers $\mathcal{T}$. The deductive closure of $\mathcal{S} \cup \mathcal{T} \cup \mathcal{G}$ under T is denoted $cl_T(\mathcal{S} \cup \mathcal{T} \cup \mathcal{G})$.

Explanations. We keep the structure of anonymous-transitive explanations, and we increase their expressive power by allowing evidence to be taken from the set $\mathcal{G}$ of gifts as well as from the set $\mathcal{T}$ of progressive transfers.

Structural results. As the generalized Lorenz dominance $\mathcal{L}$ belongs to $\mathfrak{P}_{(t,s,r,m)}$, it refines $\mathcal{P}_{(t,s,r,m)}$. $cl_T(\mathcal{S} \cup \mathcal{T} \cup \mathcal{G})$ is clearly sound and complete w.r.t. $\mathcal{P}_{(t,s,r,m)}$. $(\mathcal{T} \cup \mathcal{G})$-ATX are sound by design w.r.t. $cl_T(\mathcal{S} \cup \mathcal{T} \cup \mathcal{G})$. This nesting of preference relations collapses thanks to the following result.

Lemma 2 (Chong [17])**.** $\mathcal{L} \subseteq \bigcup_{\ell=1}^{n+1} \mathcal{E}(\mathcal{T} \cup \mathcal{G}\text{-}ATX^\ell)$

Sketch of proof. It suffices to consider a gift in first (or last) position, so as to have $\mathbf{x}^1$ and $\mathbf{y}$ with the same total income, then finding a sequence of progressive transfers from $\mathbf{x}^1$ to $\mathbf{y}$ by Lemma 1. □

Theorem 3 (Explainability of generalized Lorenz dom.)**.**

$$\bigcup_{\ell=1}^{n+1} \mathcal{E}(\mathcal{T} \cup \mathcal{G}\text{-}ATX^\ell) = cl_T(\mathcal{S} \cup \mathcal{T} \cup \mathcal{G}) = \mathcal{P}_{(t,s,r,m)} = \mathcal{L}$$

Computational aspects. From a theoretical perspective, the problem of finding short explanations for generalized Lorenz dominance statements is at least as hard as for restricted Lorenz dominance statements. Interestingly, from an algorithmic perspective, the heuristic supporting the claim of Lemma 2, i.e. systematically reducing to a problem with equal income with an initial (or final) gift is suboptimal, as evidenced by the following example.

Example 3. *Consider alternatives $\widehat{\mathbf{d}}$ and $\widehat{\mathbf{c}}$ of Example 1. Computing their cumulative sums show that $(\mathbf{d}, \mathbf{c}) \in \mathcal{L}$. The shortest explanation supporting $(\mathbf{d}, \mathbf{c})$ has length 3:*

$$\widehat{\mathbf{d}} \; \tau_{4 \to 3}^{12} \; (6\ 17\ \overline{30}\ \underline{76}\ 96) \; \mathcal{G} \; (\overline{8}\ \overline{23}\ \overline{34}\ 76\ 96) \; \tau_{5 \to 1}^{14} \; \widehat{\mathbf{c}}$$

It is strictly shorter than explanations where the gift is positioned in first or last position.

4 Explaining Robust Redistributive OWA Statements

While offering two fundamental redistributive preference models, Lorenz dominance relations remain very indecisive. In a decision-making situation [13], there might be a need for a more resolute preference structure, for example to make a choice (i.e., select a preferred alternative) or provide a ranking of alternatives. Moreover, while Lorenz dominance is inherently non-parametric, it might be useful to consider parameterized refinements capturing more specific preference patterns while still allowing for simple explanations. In turn, we shall complement the normative principles with *preference information*, both narrowing the array of possible worlds and augmenting the basis for reasoning and explanations.

[9] The *state space* is $\widehat{\mathbb{X}^n}$, the *initial, goal* and *current states* are respectively $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$ and $\mathbf{x}^k$, and the available *actions* are the self-evident comparative statements. More details are given in Section A.3

[10] In some contexts, this notion is known to be at odds with fairness [12].

	Mean length			% of ties/win/draw between expl length of methods					Mean Time (s)			% Time-out
n	HLP	**us**	opt*	HLP = **us**	HLP < **us**	HLP > **us**	**us** = opt*	**us** > opt*	HLP	**us**	opt	opt
5	3.93	3.92	3.92	99	0	1	100	0	10^{-3}	10^{-3}	.15	0
10	8.36	8.21	8.05	83	1	16	59	41	10^{-3}	10^{-3}	16.65	1.4
20	14.65	13.81	13.71	34	0	66	90	10	10^{-3}	10^{-3}	139.71	89
50	26.79	24.98	24.98	12	2	86	100	0	10^{-3}	10^{-3}	150	100

* the value of our heuristic is taken as optimum if the computation timed out (150s for one explanation).

Table 1. Comparison between our heuristic described in Fig. 1, Hardy, Littlewood and Polya [23] (HLP) and a provably optimal algorithm for $\mathcal{L}^*$

4.1 Preference based on a Set of Ordered Weighted Averaging Operators

Preferences represented by score functions that are *ordered weighted averaging* operator (OWA) originate from [41] in the context of inequality indices and [43] in the context of multiple criteria decision aiding. OWA is parameterized by a n-tuple $\boldsymbol{\omega}$ and maps an alternative $\mathbf{x}$ to $\sum_{i\in[n]} \boldsymbol{\omega}_i \widehat{\mathbf{x}}_i$, meaning the score is a ranked weighted sum. The same model is sometimes called *generalized Gini index* (GGI), as a specific value of the parameter $\boldsymbol{\omega}$ yields the Gini index. We give a definition focused on the preference structure rather than the score, and that inherently represents the skeptical inference over a set Ω of parameter values that represent incomplete preference information.

Definition 6 (robust OWA-based preference)**.** Let Ω be a non-empty subset of the L_1 unit sphere[11] of $\mathbb{R}^n$ and $\mathcal{O}_{\mathrm{WA}_\Omega}$ the binary relation defined by $(\mathbf{x},\mathbf{y}) \in \mathcal{O}_{\mathrm{WA}_\Omega}$ iff $\sum_{i\in[n]} \boldsymbol{\omega}_i \widehat{\mathbf{x}}_i \leq \sum_{i\in[n]} \boldsymbol{\omega}_i \widehat{\mathbf{y}}_i$ for all $\boldsymbol{\omega} \in \Omega$.

4.2 Properties of OWA-based Preference

We observe that given a single parameter $\boldsymbol{\omega}$, $\mathcal{O}_{\mathrm{WA}_{\{\boldsymbol{\omega}\}}}$ is:

- reflexive, transitive and symmetric whatever $\boldsymbol{\omega}$;
- monotonic when all components of $\boldsymbol{\omega}$ are non-negative, reflecting the desirability of all criteria;
- redistributive when the components of $\boldsymbol{\omega}$ are non-increasing, giving more importance to less satisfied agents.

Let[12] $\Omega^{\emptyset}$ be the set of non-negative, non-increasing, non-null vectors of the unit sphere of $\mathbb{R}^n$ i.e. the set of parameters ensuring each preference relation $\mathcal{O}_{\mathrm{WA}_{\{\boldsymbol{\omega}\}}}$ for $\omega \in \Omega^{\emptyset}$ satisfies (t), (s), (m) and (r).

Lemma 3 (Argyris et al. [2])**.** $\mathcal{L} = \mathcal{O}_{\mathrm{WA}_{\Omega^{\emptyset}}}$

This simple, yet strong result establishes that in the absence of additional preference information, a robust OWA boils down to generalized Lorenz dominance. Several characterizations of OWA-based preference have been proposed (by e.g. [41, 6, 34]). They slightly differ in the details of the properties put forward , but all of them require the relation to be *decisive* and *continuous* in some sense so as to ensure it can be represented by a real-valued function, and then impose a condition to ensure this function is additive over the set $\widehat{\mathbb{X}^n}$. We detail the result obtained by Ben-Porath and Gilboa in [6].

Property (d). $\mathcal{R}$ is *decisive* when, for all pairs of alternatives $\mathbf{x},\mathbf{y}$, either $(\mathbf{x},\mathbf{y})$ or $(\mathbf{y},\mathbf{x})$ (or both, in which case $\mathbf{x}$ and $\mathbf{y}$ are considered equally desirable) belong to $\mathcal{R}$.

$\mathcal{O}_{\mathrm{WA}_\Omega}$ is decisive iff Ω is a singleton.

Property (c). $\mathcal{R}$ is *continuous* when, for any alternative $\mathbf{z}$, the lower set $\{\mathbf{x} : (\mathbf{x},\mathbf{z}) \in \mathcal{R}\}$ and the upper set $\{\mathbf{y} : (\mathbf{z},\mathbf{y}) \in \mathcal{R}\}$ are closed.

Property (i). $\mathcal{R}$ satisfies *invariance* (w.r.t. order-preserving gifts) when, for all alternatives $\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}'$, if there is an agent $i \in [n]$ and $t \in \mathbb{R}$ s.t. $\widehat{\mathbf{x}}' = \widehat{\mathbf{x}} + t\mathbf{e}^i$ and $\widehat{\mathbf{y}}' = \widehat{\mathbf{y}} + t\mathbf{e}^i$, then $(\widehat{\mathbf{x}},\widehat{\mathbf{y}}) \in \mathcal{R} \iff (\widehat{\mathbf{x}}',\widehat{\mathbf{y}}') \in \mathcal{R}$.

Together with (t), the (i) property enforces a key feature of linear relations: the ability to reason *ceteris paribus*–i.e. everything else being equal. The preference of $\mathbf{y}$ over $\mathbf{x}$ solely depends on the acceptability of the *trade-off* $\mathbf{y}-\mathbf{x}$, regardless from the fact that it modifies $\mathbf{x}$ or some other $\mathbf{x}' \in \widehat{\mathbb{X}^n}$ (as long as $\mathbf{y}' := \mathbf{x}' + (\mathbf{y}-\mathbf{x})$ remains in $\widehat{\mathbb{X}^n}$).

Definition 7 (ceteris paribus equivalence)**.** Two comparative statements $(\mathbf{x},\mathbf{y})$ and $(\mathbf{x}',\mathbf{y}')$ are equivalent ceteris paribus when the alternatives $\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}'$ all belong to $\widehat{\mathbb{X}^n}$ and the vectors $\mathbf{x}-\mathbf{y}$ and $\mathbf{x}'-\mathbf{y}'$ are equal. In this case, $(\mathbf{x},\mathbf{y})$ and $(\mathbf{x}',\mathbf{y}')$ represent the same *trade-off*.

When a preference relation $\mathcal{R}$ satisfies (t) and (i), it is compatible to ceteris paribus equivalence, in the sense that two equivalent statements are either both in $\mathcal{R}$, or neither in $\mathcal{R}$. Thus, $\mathcal{R}$ can be defined by its set of *acceptable trade-offs* $to(\mathcal{R}) := \{\widehat{\mathbf{x}}-\widehat{\mathbf{y}}, (\widehat{\mathbf{x}},\widehat{\mathbf{y}}) \in \mathcal{R}\}$.

Lemma 4 (Ben-Porath and Gilboa [6])**.** *A reflexive binary relation $\mathcal{R}$ satisfies (t), (s), (r), (m), (d), (c) and (i) iff there is a n-tuple $\boldsymbol{\omega}$ of non-negative, non-increasing real numbers such that $\mathcal{R} = \mathcal{O}_{\mathrm{WA}_{\{\boldsymbol{\omega}\}}}$.*

This is a powerful representation theorem but, as we shall see, the properties put forward prove cumbersome in the light of our agenda of representing the underlying reasoning with deductive rules.

4.3 Beyond Decisiveness

Property (d) is unsatisfying for two reasons. From a reasoning standpoint, it amounts to introducing the *excluded middle* rule into the formal system, which allows to use proofs by refutation[13] as a powerful deductive tool. However, in some contexts, it might not align well with our requirement of intelligible explanations, and intuitionistic logic offers an alternative path. From a preference representation standpoint, decisiveness conflicts with the need to catch the inevitable incompleteness of information. As a clear indication of the fragility of deductions made under this requirement, note that it is the only property mentioned in this paper that is not stable under intersection. In order to smoothly transition from Lorenz-dominance (obtained with $\Omega = \Omega^{\emptyset}$) to decisiveness (obtained when Ω is a singleton), we adopt the paradigm of *robust preference learning* [22, 21]. We suppose we are given some *preference information* $\mathcal{I}$ in the form

[11] This condition ensures non-trivialness of the relation (as a null parameter corresponds to complete indifference) and the non-redundancy of the parameters (as a linear transform of the parameter yields the same OWA relation), while the choice of the L_1 norm ensures tractability.

[12] This notation is consistent with Def. 8.

[13] Indeed, to prove that $\mathbf{x}$ is preferred to $\mathbf{y}$, it would suffice to prove that $\mathbf{y}$ cannot be preferred to $\mathbf{x}$.

of a binary relation over alternatives, i.e. a set of reference comparative statements that are required to hold in the sought preference relation – $\mathcal{I}$ can thus be seen as a curated learning set.

Property (pi$^{\mathcal{I}}$). With $\mathcal{I}$ a binary relation over alternatives, $\mathcal{R}$ is compatible to *preference information* $\mathcal{I}$ when $\mathcal{R} \supseteq \mathcal{I}$.

Deductively, compatibility to preference information is straightforwardly represented by considering statements of $\mathcal{I}$ as self-evident. In terms of numeric representation, we need to consider the set $\Omega^{\mathcal{I}}$ containing exactly all the parameters such that $\mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$ is compatible to $\mathcal{I}$. Of course, we have to assume this requirement is feasible.

Definition 8 (compatible parameter set)**.** With $\mathcal{I}$ a binary relation over alternatives, let $\Omega^{\mathcal{I}}$ the set of non-negative, non-increasing vectors $\boldsymbol{\omega}$ of the unit sphere of $\mathbb{R}^n$ s.t. $\mathcal{O}_{\text{WA}_{\{\boldsymbol{\omega}\}}} \supseteq \mathcal{I}$. When $\Omega^{\mathcal{I}} \neq \emptyset$, $\mathcal{I}$ is said *consistent* (with the fair OWA model).

While skeptic inference often face computational issues [19], it remains in our case polynomial, as OWA operator is linear w.r.t. its parameters.

Lemma 5 (inspired by [22])**.** *Given a binary relation over alternatives $\mathcal{I}$, the set $\Omega^{\mathcal{I}}$ is the polytope of $\mathbb{R}^n$ defined by the linear constraints over the variable $\boldsymbol{\omega}$: $\boldsymbol{\omega}_i \geq 0$ for all $i \in [n]$; $\boldsymbol{\omega}_i - \boldsymbol{\omega}_{i+1} \geq 0$ for all $i \in [n-1]$, $\sum_{i\in[n]} \boldsymbol{\omega}_i = 1$ and $\sum_{i\in[n]} (\widehat{\mathbf{b}}_i - \widehat{\mathbf{a}}_i)\boldsymbol{\omega}_i \geq 0$ for all $(\mathbf{a}, \mathbf{b}) \in \mathcal{I}$. Hence, checking if $\mathcal{I}$ is consistent, or if a given comparative statement belongs to $\mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$ reduce to linear optimization problems that can be solved in polynomial time.*

Example 4. *Consider the preference information $(\mathbf{b}, \mathbf{d})$ given by the central authority in Example 1. The OWA score parameterized by any $\boldsymbol{\omega} \in \Omega^{\mathcal{I}}$ to the alternative $\mathbf{d}$ should be greater than the one of $\mathbf{b}$. Hence $\widehat{\mathbf{d}} \cdot \boldsymbol{\omega} \geq \widehat{\mathbf{b}} \cdot \boldsymbol{\omega}$, or equivalently $(\widehat{\mathbf{d}} - \widehat{\mathbf{b}}) \cdot \boldsymbol{\omega} \geq 0$. This constraint can be interpreted as the trade-off $(\widehat{\mathbf{d}} - \widehat{\mathbf{b}}) = (+4, -8, -10, +4, -2)$ being desirable.*

4.4 A Characterization of Robust OWA Preference

Our next goal is to characterise preference structures induced by sets of OWA through actionable properties allowing to build a sound and complete explanation engine. We would like to keep (t), (s), (r), (m), and (i) but drop (d), (c). We have just given reasons for the former, which we would like to replace by (pi$^{\mathcal{I}}$), and while the latter is a fantastic mathematical tool, it is a cognitive nightmare. Continuity essentially allows taking the limit on the left and on the right in a sequence of preference statements, but begs the questions of defining sequences, checking their convergence and computing their limit. One can hardly imagine a non-expert understanding and discussing such a property.

However, as we shall see below, (t), (s), (r), (m), (c), (i) and (pi$^{\mathcal{I}}$) are not sufficient to fully characterize robust fair OWA. This is why we now introduce a new notion, close to the ceteris paribus equivalence, but slightly stronger.

Starting from the ceteris paribus equivalence relation, we consider incorporating symmetry and relaxing the vector equality between trade-offs into the existence of a non-negative link, ending up in a stronger property defined below.

Definition 9 (congruent comparative statements)**.** Two comparative statements $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$ are congruent when $\widehat{\mathbf{x}} - \widehat{\mathbf{y}}$ and $\widehat{\mathbf{x}}' - \widehat{\mathbf{y}}'$ are non-negatively linked[14]. In this case, we write $(\mathbf{x}, \mathbf{y}) \equiv (\mathbf{x}', \mathbf{y}')$.

[14] I.e. at least one of the vectors is null, or there is $\lambda > 0$ s.t. $(\widehat{\mathbf{x}} - \widehat{\mathbf{y}}) = \lambda(\widehat{\mathbf{x}}' - \widehat{\mathbf{y}}')$.

Property (cc). $\mathcal{R}$ is *compatible to congruence* when, for all alternatives $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$, if $(\mathbf{x}, \mathbf{y}) \equiv (\mathbf{x}', \mathbf{y}')$ then $(\mathbf{x}, \mathbf{y}) \in \mathcal{R} \iff (\mathbf{x}', \mathbf{y}') \in \mathcal{R}$.

Note that the property (cc) entails (i) and (s). It is instrumental in the characterization of decisive OWA preference[15]. One can wonder if this new structure of acceptable trade-offs can be derived, or logically follows, from the properties (t), (s), (r), (m), (c), (i) and (pi$^{\mathcal{I}}$). The answer to both questions is negative.

Theorem 4. *When $\Omega^{\mathcal{I}}$ is not a singleton, property (cc) cannot be deduced from (t), (s), (r), (m), (c), (i) and (pi$^{\mathcal{I}}$).*

Counter-example. Let $\Gamma := \{(z_1, z_2, z_3) \in \mathbb{R}^3$ satisfying $(1)\ 3z_1 + 2z_2 + z_3 \geq 3$ or $(2)\ z_1 \geq 0$ and $z_1 + z_2 \geq 0$ and $z_1 + z_2 + z_3 \geq 0\}$, and $\mathcal{R}$ the binary relation over $\mathbb{R}^3$ defined by $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ iff $\widehat{\mathbf{y}} - \widehat{\mathbf{x}} \in \Gamma$. $\mathcal{R}$ satisfies (i) and (s) by construction. It is continuous because the set Γ of its acceptable trade-offs is closed (as the union of intersections of preimages of closed sets by continuous functions). It is transitive because Γ is stable under addition (the sum of two vectors satisfying (1) or (2) respectively satisfies (1) or (2), and the sum of one satisfying (1) and the other (2) satisfies (2)). (r) and (m) are enforced with condition (2). Nevertheless, while the trade-off $t_1 := (2, -4, 6)$ is acceptable (corresponding e.g. to the comparative statement $(0, 8, 10)$ vs $(2, 4, 16)$), the trade off $\frac{1}{2}t_1 = (1, -2, 3)$ is not (while it corresponds e.g. to the comparative statement $(3, 8, 9)$ vs $(4, 6, 12)$). □

As a corollary[16], properties (t), (s), (r), (m), (c), (i) and (pi$^{\mathcal{I}}$) are not sufficient to characterize the robust OWA preference relation $\mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$. Introducing the inference rule

$$\textbf{Rule CC :}\ \frac{(\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{d}) \equiv (\mathbf{a}, \mathbf{b})}{(\mathbf{c}, \mathbf{d})}\ \text{(compatibility to congruence)}$$

corresponding to property (cc) we are now ready to introduce one of our main results.

Theorem 5. *For any preference information $\mathcal{I}$ which is consistent with the OWA model:*

$$cl_{T,CC}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I}) = \mathcal{P}_{(t,r,m,cc,pi^{\mathcal{I}})} = \mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$$

Theorem 5 characterizes semantically and deductively robust OWA-based preference: $\mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$ is the only reflexive binary relation satisfying (t), (s), (r), (m), (cc) and (pi$^{\mathcal{I}}$); moreover $\mathbf{x}$ is less preferred than $\mathbf{y}$ according to this relation if, and only if, there is a proof establishing this statement using the deductive rules T and CC starting from self-evident statements in $\mathcal{T}$, $\mathcal{G}$ or $\mathcal{I}$. As noted in Section 3 in the case of Lorenz dominance, the chain of inclusion from left to right denoting *soundness*, i.e. $cl_{T,CC}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I}) \subseteq \mathcal{P}_{(t,r,m,cc,pi^{\mathcal{I}})} \subseteq \mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$ is structurallly valid because the normative properties put forward match both those of the robust OWA and the chosen derivation rules. We will now devise an explanation engine implementing a restriction of $cl_{T,CC}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I})$ and *complete* w.r.t. $\mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$, that will close the proof of Th. 5.

[15] A key step of the proof establishes the equation when the link coefficient λ is the reciprocal of a positive integer, reasoning with transitivity and the excluded middle.

[16] The counter-example arbitrarily focuses on the family of trade-offs $\{\delta : \boldsymbol{\omega} \cdot \delta \geq K\}$, with $\boldsymbol{\omega} = (3, 2, 1)$, but can be altered to incorporate any consistent but non-decisive preference information.

4.5 Explanation Schemes

Example 1 illustrates a case where a comparative statement is supported by an ATX of length 3 with self-evident statements from $\mathcal{S}$, $\mathcal{G}$, $\mathcal{T}$ and also from $cl_{CC}(\mathcal{I})$ –statements that are congruent to one appearing in the preference information. Nevertheless, this explanation scheme might not be complete w.r.t. $\mathcal{O}_{\text{WA}_{\Omega^\mathcal{I}}}$, and is certainly cumbersome computationally, because the constraint of remaining inside $\widehat{\mathbb{X}^n}$ is difficult to satisfy, especially when both alternatives of the claimed statement are close to the border. We thus propose to relax the requirement of finding a path from $\widehat{\mathbf{x}}$ to $\widehat{\mathbf{y}}$ into finding a path from $\mathbf{x}'$ to $\mathbf{y}'$ with $(\mathbf{x}', \mathbf{y}')$ congruent to $(\mathbf{x}, \mathbf{y})$.

Definition 10 (CTX scheme)**.** Let $\mathcal{R}$ be a binary relation over $\widehat{\mathbb{X}^n}$. A *congruent-transitive explanation based on evidence in* $\mathcal{R}$ ($\mathcal{R}$-CTX) of length ℓ is a pair (s, c) where the *support* s is a ℓ-tuple of comparative statements $s = ((\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^\ell, \mathbf{y}^\ell)) \in cl_{CC}(\mathcal{R})^\ell$ and the *claim* c is a comparative statement $c = (\mathbf{x}, \mathbf{y})$ satisfying $(\mathbf{x}^1, \mathbf{y}^\ell) \equiv (\mathbf{x}, \mathbf{y})$ and, for all integers k between 2 and ℓ, $\mathbf{x}^k = \mathbf{y}^{k-1}$.

As it only uses deductive rules and basic truths, the CTX scheme is sound w.r.t the deductive rules T and CC. As another main result, we establish its completeness by a constructive proof from which can be extracted a tractable algorithm. It concludes the proof of Th. 5.

Theorem 6. $\mathcal{O}_{\text{WA}_{\Omega^\mathcal{I}}} \subseteq \bigcup_{\ell=1}^{n+|\mathcal{I}|} \mathcal{E}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I}\text{-}CTX^\ell)$

Proof. Let $(\mathbf{x}, \mathbf{y}) \in \mathcal{O}_{\text{WA}_{\Omega^\mathcal{I}}}$. By Farkas' lemma applied to the MILO formulation of Lemma 5, there are non-negative coefficients $\langle \lambda^*_{(\mathbf{a},\mathbf{b})} \rangle_{(\mathbf{a},\mathbf{b})\in\mathcal{I}}$, $\langle \mu^*_i \rangle_{i\in[n]}$ and $\langle \nu^*_{j,i} \rangle_{1\le i<j\le n}$ s.t.

$$\widehat{\mathbf{y}} - \widehat{\mathbf{x}} = \sum_{(\mathbf{a},\mathbf{b})\in\mathcal{I}} \lambda^*_{(\mathbf{a},\mathbf{b})}(\widehat{\mathbf{b}} - \widehat{\mathbf{a}}) + \sum_{i\in[n]} \mu^*_i \mathbf{e}^i + \sum_{i<j} \nu^*_{j,i}(\mathbf{e}^i - \mathbf{e}^j)$$

This equation additively decomposes the trade-off corresponding to the comparative statement into 3 terms, each one a summation of trade-offs corresponding to self-evident statements respectively in $cl_{CC}(\mathcal{I})$, $\mathcal{G}$ and $\mathcal{T}$. For every agent i, we split this equation into two parts, one containing the sum of positive values Δ_i^+ and the other containing the sum of negative values Δ_i^-, such that $\widehat{\mathbf{y}}_i - \widehat{\mathbf{x}}_i = \Delta_i^+ + \Delta_i^-$. If we define $\widehat{\mathbf{x}}'$ and $\widehat{\mathbf{y}}'$ such that $\forall i \in [n]$: $\widehat{\mathbf{x}}'_i = \widehat{\mathbf{x}}_i + \sum_{j=1}^{i} \Delta_{j-1}^+ - \sum_{j=1}^{i} \Delta_j^-$ and $\widehat{\mathbf{y}}'_i = \widehat{\mathbf{y}}_i + \sum_{j=1}^{i} \Delta_{j-1}^+ - \sum_{j=1}^{i} \Delta_j^-$. One can check that *(i)* $\widehat{\mathbf{y}}_i - \widehat{\mathbf{x}}_i = \widehat{\mathbf{y}}'_i - \widehat{\mathbf{x}}'_i$, hence $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$ are congruent; and *(ii)* $\forall i \in [n]$ $\widehat{\mathbf{x}}'_i - \widehat{\mathbf{x}}'_{i-1} \ge \Delta_{i-1}^+ - \Delta_i^-$, hence the separation between criteria allows $\widehat{\mathbf{x}}'_{i-1}$ to be increased and $\widehat{\mathbf{x}}'_i$ to be decreased by the values given by the Farkas certificate while keeping $\widehat{\mathbf{x}}'_{i-1} \le \widehat{\mathbf{x}}'_i$, thus to perform in any order the $\mathcal{T}$, $\mathcal{G}$ and $cl_{CC}(\mathcal{I})$ statements described by $\langle \lambda^*_{(\mathbf{a},\mathbf{b})} \rangle_{(\mathbf{a},\mathbf{b})\in\mathcal{I}}$, $\langle \mu^*_i \rangle_{i\in[n]}$ and $\langle \nu^*_{j,i} \rangle_{1\le i<j\le n}$ to build the transitive chain between $\mathbf{x}'$ and $\mathbf{y}'$. It results that we can always provide a $(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I})$-CTX of length at most $|\mathcal{I}| + n$. □

The use of a certificate of infeasibility to derive an explanation can also be found in [24, 4]. Obtaining such a certificate with a strong duality result (here, Farkas' lemma) can also be found in [26, 3, 5].

Example 5. *We propose to compute a CTX as explained in the proof of Theorem 6 for the statement* $(\mathbf{a}, \mathbf{c})$*, with* $\widehat{\mathbf{c}} = (22\ 23\ 34\ 76\ 82)$ *over* $\widehat{\mathbf{a}} = (16\ 31\ 51\ 70\ 83)$*. The Farkas certificate obtained is:* $\lambda^*_{(\mathbf{b},\mathbf{d})} = 1.5$*,* $\mu^* = (0\ 0\ 2\ 0\ 2)$ *and a single redistributive transfer* $\nu^*_{3,2} = 4$*. We can assess its validity by computing* $\widehat{\mathbf{c}} - \widehat{\mathbf{a}} = (+6, -8, -17, +6, -1) = 1.5 * (+4, -8, -10, +4, -2) + (0, 0, +2, 0, +2) + (0, +4, -4, 0, 0)$*. We compute the alternatives* $\mathbf{x}'$ *and* $\mathbf{y}'$ *as described in the proof. For agent 1, we have* $\Delta_1^+ = 6$ *and* $\Delta_1^- = 0$*, hence* $\mathbf{y}'_1 = \widehat{\mathbf{c}}_1 - \Delta_1^- = \widehat{\mathbf{c}}_1 = 22$ *and* $\mathbf{x}'_1 = \widehat{\mathbf{a}}_1 - \Delta_1^- = \widehat{\mathbf{a}}_1 = 16$*. For agent 2 we have* $\Delta_2^+ = 4$ *(from the redistributive transfer) and* $\Delta_2^- = -12$ *(from the* $\mathcal{I}$*-congruent statement), hence* $\mathbf{y}'_2 = \widehat{\mathbf{c}}_2 + \Delta_1^+ - \Delta_2^- - \Delta_1^- = 23 + 6 - (-12) = 41$ *and* $\mathbf{x}'_2 = \widehat{\mathbf{a}}_2 + \Delta_1^+ - \Delta_2^- - \Delta_1^- = 31 + 6 - (-12) = 49$*. For agent 3, we have* $\Delta_3^+ = -15 - 4 = -19$ *(from the redistributive transfer and the* $\mathcal{I}$*-congruent statement) and* $\Delta_3^- = 20$ *(from the gift), hence* $\mathbf{y}'_3 = \widehat{\mathbf{c}}_3 + \Delta_2^+ + \Delta_1^+ - \Delta_3^- - \Delta_2^- - \Delta_1^- = 34 + 4 + 6 - (-19) - (-12) = 75$ *and* $\mathbf{x}'_3 = \widehat{\mathbf{a}}_3 + \Delta_2^+ + \Delta_1^+ - \Delta_3^- - \Delta_2^- - \Delta_1^- = 51 + 4 + 6 - (-19) - (-12) = 92$*. We continue for agents 4 and 5 and obtain* $\mathbf{y}' = (22\ 41\ 75\ 119\ 134)$ *and* $\mathbf{x}' = (16\ 49\ 92\ 113\ 135)$*. We have* $\mathbf{y}' - \mathbf{x}' = \widehat{\mathbf{c}} - \widehat{\mathbf{a}}$*, the two pairs are congruent, we can then build the CTX of length 3 from the chain of alternatives* $(\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4)$ *with* $\mathbf{x}^1 := \mathbf{x}'$ *and* $\mathbf{x}^4 := \mathbf{y}'$ *as CTX. Any permutation of the redistributive transfer, gift and PI-congruent statement are possible, if we opt for this ordering we have* $\mathbf{x}^2 := (16\ 53\ 88\ 113\ 135)$ *and* $\mathbf{x}^3 := (22\ 41\ 75\ 119\ 134)$*.*

One can wonder if an ATX, rather a CTX, can be found to support a claim in $\mathcal{O}_{\text{WA}_{\Omega^\mathcal{I}}}$. This is possible under mild conditions: let $\mathbb{F} := \{\mathbf{z} \in \widehat{\mathbb{X}^n} : \exists i\ \mathbf{z}_i = 0 \text{ or } \exists j \neq i\ \mathbf{z}_i = \mathbf{z}_j\}$. $\mathbb{F}$ is the frontier of $\widehat{\mathbb{X}^n}$.

Theorem 7. *Let* $\mathcal{I}$ *be a binary relation over alternatives consistent with the OWA model and* $(\mathbf{x}, \mathbf{y}) \in \mathcal{O}_{\text{WA}_{\Omega^\mathcal{I}}}$*. If:*

i. neither alternatives $\mathbf{x}$ *and* $\mathbf{y}$ *belong to* $\mathbb{F}$*; or*
ii. $\sup_{\boldsymbol{\omega}\in\Omega^\mathcal{I}} \boldsymbol{\omega} \cdot (\widehat{\mathbf{y}} - \widehat{\mathbf{x}}) > 0$

then there is a $[\mathcal{T} \cup \mathcal{G} \cup cl_{CC}(\mathcal{I})]$*-ATX supporting* $(\mathbf{x}, \mathbf{y})$*.*

As our proof relies on the construction of an ATX of unbounded length, we conjecture the existence of comparative statements not supported by an ATX.

5 Discussion and Perspectives

We derived sound and complete explanations schemes for robust OWAs (a.k.a. Generalized Gini Index), using a formal model based on logic. We are not the first to follow such a path, and our approach is akin to the ones initiated by [15] and expanded by [35, 10, 32, 11]. We ensure that our explanation are narratively and cognitively friendly by avoiding the use of purely technical properties (such as continuity), and by providing step-wise, chained explanations of bounded length [9].

We think our results fill an important gap, as OWAs are commonly used to account for fairness in combinatorial optimisation [33, 27], computational social choice [2, 28], preference or reinforcement learning [14, 18]. It is reasonable to assume that, whenever fairness is important, we also want to be able to scrutinize the obtained decisions and avoid as much as unwarranted biases or instabilities due to the choice of specific parameters. By providing provably correct and readable explanations [17], we answer this need. Our explanation could indeed enable the assessment of the procedural regularity and general adequacy of algorithmic decisions, or even recourse [25, 16].

Regarding trustworthy AI, the next item in our agenda would be to submit explanation to *probation* of the underlying requirements, with *critical questions* [40] such as "is it reasonable to be symmetric?" (maybe one of the agents deserves a privileged treatment) or "are we sure the utilities use a common scale?". This should require to embed the explanation engine inside a dialectical agent capable of non-monotonic reasoning. Regarding decision theory, the next step is to *transfer* the proposed approach to complex models such as Choquet integrals, yet this may require heavy axiomatic work.

[17] Of the scientific rather than everyday sort [37], hence maybe not totally in-line with expected canon of the latter [31].

References

[1] S. Angilella, S. Greco, and B. Matarazzo. Non-additive robust ordinal regression: A multiple criteria decision model based on the Choquet integral. *European Journal of Operational Research*, 201(1):277–288, 2010. Number: 1 Publisher: Elsevier.

[2] N. Argyris, Ö. Karsu, and M. Yavuz. Fair resource allocation: Using welfare-based dominance constraints. *European journal of operational research*, 297(2):560–578, 2022.

[3] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183, 2017. Number: 2 Publisher: Springer.

[4] K. Belahcène, Y. Chevaleyre, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Accountable approval sorting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 70–76. ijcai.org, 2018.

[5] K. Belahcène, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Comparing options with argument schemes powered by cancellation. In S. Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1537–1543, Macao, China, 2019. ijcai.org.

[6] E. Ben-Porath and I. Gilboa. Linear measures, the gini index and the income-equality tradeoff. *Journal of Economic Theory*, 64:443–467, 1994.

[7] N. Benabbou, C. Gonzales, P. Perny, and P. Viappiani. Minimax regret approaches for preference elicitation with rank-dependent aggregators. *EURO journal on Decision processes*, 3:29–64, 2015.

[8] N. Benabbou, C. Leroy, T. Lust, and P. Perny. Combining local search and elicitation for multi-objective combinatorial optimization. In *Algorithmic Decision Theory: 6th International Conference, ADT 2019*, pages 1–16, 2019.

[9] I. Bleukx, J. Devriendt, E. Gamba, B. Bogaerts, and T. Guns. Simplifying Step-Wise Explanation Sequences. In *29th International Conference on Principles and Practice of Constraint Programming (CP 2023)*, volume 280 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. ISBN 978-3-95977-300-3.

[10] A. Boixel, U. Endriss, and R. de Haan. A calculus for computing structured justifications for election outcomes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 4859–4866. AAAI Press, 2022.

[11] A. Boixel, U. Endriss, and O. Nardi. Displaying justifications for collective decisions. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 5892–5895. ijcai.org, 2022.

[12] S. Bouveret, Y. Chevaleyre, and N. Maudet. Fair allocation of indivisible goods. In F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, editors, *Handbook of Computational Social Choice*, pages 284–310. Cambridge University Press, 2016.

[13] 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, 1st edition, 2006. ISBN 0-387-31098-3.

[14] R. Busa-Fekete, B. Szörényi, P. Weng, and S. Mannor. Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, pages 625–634. PMLR, 2017.

[15] O. Cailloux and U. Endriss. Arguing about voting rules. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, pages 287–295. ACM, 2016.

[16] R. Chatila, V. Dignum, M. Fisher, F. Giannotti, K. Morik, S. Russell, and K. Yeung. Trustworthy AI. In B. Braunschweig and M. Ghallab, editors, *Reflections on Artificial Intelligence for Humanity*, volume 12600 of *Lecture Notes in Computer Science*, pages 13–39. Springer, 2021.

[17] K.-M. Chong. Doubly stochastic operators and rearrangement theorems. *Journal of Mathematical Analysis and Applications*, 56(2):309–316, 1976.

[18] V. Do, S. Corbett-Davies, J. Atif, and N. Usunier. Two-sided fairness in rankings via lorenz dominance. *Advances in Neural Information Processing Systems*, 34:8596–8608, 2021.

[19] T. Eiter and G. Gottlob. On the complexity of propositional knowledge base revision, updates, and counterfactuals. *Artificial Intelligence*, 57 (2-3), 1992.

[20] M. R. Garey and D. S. Johnson. Computers and intractability: a guide to the theory of np-hardness, 1979.

[21] S. Greco, B. Matarazzo, and R. Slowinski. Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129(1):1–47, 2001. Number: 1.

[22] S. Greco, V. Mousseau, and R. Słowiński. Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191(2):416–436, 2008. Number: 2.

[23] G. Hardy, J. Littlewood, and G. Pólya. Some simple inequalities satisfied by convex functions. *Messenger of Mathematics*, 58:145–152, 1929.

[24] U. Junker. QUICKXPLAIN: preferred explanations and relaxations for over-constrained problems. In D. L. McGuinness and G. Ferguson, editors, *AAAI 2004*, pages 167–172, 2004.

[25] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165(3):633–705, 2017. ISSN 00419907.

[26] C. Labreuche, N. Maudet, and W. Ouerdane. Justifying dominating options when preferential information is incomplete. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 486–491, Amsterdam, Netherlands, 2012. IOS Press.

[27] J. Lesca, M. Minoux, and P. Perny. The fair owa one-to-one assignment problem: Np-hardness and polynomial time special cases. *algorithmica*, 81:98–123, 2019.

[28] J. W. Lian, N. Mattei, R. Noble, and T. Walsh. The conference paper assignment problem: Using order weighted averages to assign indivisible goods. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[29] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[30] R. Marinescu, D. Bhattacharjya, J. Lee, F. Cozman, and A. Gray. Credal marginal map. In *Annual Conference on Neural Information Processing Systems*, 2023.

[31] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

[32] O. Nardi, A. Boixel, and U. Endriss. A graph-based algorithm for the automated justification of collective decisions. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, pages 935–943, 2022.

[33] W. Ogryczak. Fair optimization–methodological foundations of fairness in network resource allocation. In *2014 IEEE 38th International Computer Software and Applications Conference Workshops*, pages 43–48. IEEE, 2014.

[34] P. Perny, O. Spanjaard, and L. Storme. A decision-theoretic approach to robust optimization in multivalued graphs. *Ann. Oper. Res.*, 147(1): 317–341, 2006.

[35] D. Peters, A. D. Procaccia, A. Psomas, and Z. Zhou. Explainable voting. In *Advances in Neural Information Processing Systems NeurIPS 2020*.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[37] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Häb-Umbach, I. Horwath, E. Hüllermeier, F. Kern, S. Kopp, K. Thommes, A. N. Ngomo, C. Schulte, H. Wachsmuth, P. Wagner, and B. Wrede. Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Trans. Cogn. Dev. Syst.*, 13(3): 717–728, 2021.

[38] T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Coherent choice functions under uncertainty. *Synthese*, 172:157–176, 2010.

[39] A. F. Shorrocks. Ranking Income Distributions. *Economica*, 50(197): 3–17, 1983.

[40] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, Cambridge, England, 2008.

[41] J. A. Weymark. Generalized gini inequality indices. *Mathematical Social Sciences*, 1(4):409–430, 1981.

[42] H. Willot, K. Belahcene, and S. Destercke. Source code, 2024. URL https://github.com/BOB-Henoik/ECAI2024-Principled_Explanations_for_Robust_Redistributive_Decisions.

[43] R. R. Yager. On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking. In D. Dubois, H. Prade, and R. R. Yager, editors, *Readings in Fuzzy Sets for Intelligent Systems*, pages 80–87. Morgan Kaufmann, 1993.

A Appendix

A.1 Proof of Theorem 1

We prove the "soundness" inclusions:

$$\bigcup_{\ell=1}^{+\infty} \mathcal{E}(\mathcal{T}\text{-ATX}^{\ell}) \subseteq cl_T(\mathcal{S} \cup \mathcal{T}) \subseteq \mathcal{P}_{(t,s,r)} \subseteq \mathcal{L}^*$$

- $\mathcal{P}_{(t,s,r)} \subseteq \mathcal{L}^*$: it suffices to show that $\mathcal{L}^*$ satisfies the properties of reflexivity (obvious), (t), (s) and (r), thus belongs to $\mathfrak{P}_{(t,s,r)}$ and refines $\mathcal{P}_{(t,s,r)}$. Let $\Lambda_i(\mathbf{a}) := \sum_{k=1}^{i} \widehat{\mathbf{a}}_k$ denote the i^{th} component of the Lorenz vector of $\mathbf{a}$.
  - Transitivity follows from the numerical representation of preference. Each Lorenz component is a score function and induces a total preorder. The intersection of these preorders is itself a preorder.
  - Symmetry is a consequence of the definition of the Lorenz components from the sorted components of an alternative. Hence, all the permutations of an alternative map to the same Lorenz vector.
  - Let us examine the influence of a redistributive transfer on the Lorenz components of an alternative. Suppose $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}$ and i, j, t s.t. $(\mathbf{x}, \mathbf{y}) = \tau_{j \to i}^{t}$. For any $i^* \in [n]$,

$$\Lambda_{i^*}(\mathbf{y}) = \sum_{k=1}^{i^*} \widehat{\mathbf{y}}_k = \begin{cases} \Lambda_{i^*}(\mathbf{x}), & \text{if } i^* < i \\ \Lambda_{i^*}(\mathbf{x}) + t \geq \Lambda_{i^*}(\mathbf{x}), & \text{if } i \leq i^* \\ & \quad \& \ i^* < j \\ \Lambda_{i^*}(\mathbf{x}) + t - t = \Lambda_{i^*}(\mathbf{x}), & \text{if } i^* \geq j. \end{cases}$$

    hence $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}^*$
- $cl_T(\mathcal{S} \cup \mathcal{T}) \subseteq \mathcal{P}_{(t,s,r)}$: we need to prove that every statement that can be proved holds in every possible world. Indeed, every relation in $\mathfrak{P}_{(t,s,r)}$ contains the basic truths $\mathcal{S} \cup \mathcal{T}$ and their transitive closure, thus $cl_T(\mathcal{S} \cup \mathcal{T})$ is in their intersection.
- $\bigcup_{\ell=1}^{+\infty} \mathcal{E}(\mathcal{T}\text{-ATX}^{\ell}) \subseteq cl_T(\mathcal{S} \cup \mathcal{T})$: an ATX supporting the conclusion $(\mathbf{x}, \mathbf{y})$ can be seen as a transitive chain between $\mathbf{x}$ and $\mathbf{y}$, where the first (i.e. $(\mathbf{x}, \widehat{\mathbf{x}})$) and last (i.e. $(\widehat{\mathbf{y}}, \mathbf{y})$) comparative statements are in $\mathcal{S}$, and every other statement is in $\mathcal{T}$.

Moreover, the "completeness" result of Lemma 1 ensures $\mathcal{L}^* \subseteq \bigcup_{\ell=1}^{n} \mathcal{E}(\mathcal{T}\text{-ATX}^{\ell})$, hence every inclusion collapses to equality (and the size of explanations is bounded by n).

A.2 Proof of Theorem 2

Hardness is obtained by reduction from 3-Partition [20]. The input of the 3-Partition problem is a set S of integers such that $|S| = 3m$ and $\sum_{s \in S} s = mT$. The goal is to answer the problem "Can we build a partition of S in m subsets of size exactly 3 where the sum of values inside each subset is exactly T ?". Without loss of generality we assume $\frac{T}{4} < s < \frac{T}{2} \ \forall s \in S$.

Let $\mathbf{x}$ and $\mathbf{y}$ be two alternatives from $\widehat{\mathbb{X}^n}$[18] with $n = 4m$ defined by :

$$\mathbf{x}_i = \begin{cases} \sum_{j=1}^{i-1} \widehat{S_j} & \text{if } 1 \leq i \leq 3m \\ iT & \text{if } 3m < i \leq 4m \end{cases}$$

$$\mathbf{y}_i = \begin{cases} \sum_{j=1}^{i} \widehat{S_j} & \text{if } 1 \leq i \leq 3m \\ (i-1)T & \text{if } 3m < i \leq 4m \end{cases}$$

[18] we directly define $\mathbf{x}$ and $\mathbf{y}$ reordered to avoid the $\widehat{\cdot}$ notation.

Thus,

$$\mathbf{y}_i - \mathbf{x}_i = \begin{cases} \widehat{S_i} & \text{if } 1 \leq i \leq 3m \\ -T & \text{if } 3m < i \leq 4m \end{cases}$$

Let $pros(\mathbf{a}, \mathbf{b}) := \{i \in [n] : \mathbf{a}_i < \mathbf{b}_i\}$, such that $pros(\mathbf{x}, \mathbf{y}) = [3m]$ and $pros(\mathbf{y}, \mathbf{x}) = [4m] \setminus [3m]$, i.e. indices between 1 and $3m$ are pro-$\mathbf{y}$ while indices between $3m+1$ and $4m$ are pro-$\mathbf{x}$.

As $\mathbf{x}, \mathbf{y} \in \widehat{\mathbb{X}^n}$, $\Lambda(\mathbf{y}) - \Lambda(\mathbf{x}) = \Lambda(\mathbf{y} - \mathbf{x})$:

$$\Lambda_i(\mathbf{y}) - \Lambda_i(\mathbf{x}) = \begin{cases} \sum_{j=1}^{i} \widehat{S_j} & \text{if } 1 \leq i \leq 3m \\ \sum_{j=1}^{3m} \widehat{S_j} - (i-3m)T & \text{if } 3m < i \leq 4m \\ = mT - (i-3m)T & \end{cases}$$

The difference between the i-th component of the Lorenz vectors of $\mathbf{y}$ and $\mathbf{x}$ is strictly positive for $1 \leq i < 4m$ and null for $i = 4m$, thus $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}^*$.

If S is a positive instance of the 3-Partition problem, we show the claim $(\mathbf{x}, \mathbf{y})$ is supported by a $\mathcal{T}$-ATX of length $3m$. Let $((S_{i_1^1}, S_{i_2^1}, S_{i_3^1}), \ldots, (S_{i_1^m}, S_{i_2^m}, S_{i_3^m}))$ denote the partition of S into triplets of sum T. We build two sequences of alternatives $\mathbf{x}^1, \ldots, \mathbf{x}^{3m}$ and $\mathbf{y}^1, \ldots, \mathbf{y}^{3m}$ the following way: $\mathbf{x}^1 := \mathbf{x}$, and at each step k, $\mathbf{y}^k$ is deduced from $\mathbf{x}^k$ by transferring to agent $i := 3m + 1 - k$ (i.e. traversing the receiving agents in $[3m]$ from most satisfied to least satisfied) the amount $\widehat{S}_i$ from the agent j such that $\widehat{S}_i$ belongs to the j-th triplet. Finally $\mathbf{x}^{k+1} := \mathbf{y}^k$. The proofs that $\mathbf{x}^k$ remains sorted and that $\mathbf{y}^{3m} = \mathbf{y}$ are left to the reader.

Reciprocally, if the comparative statement (x, y) is supported by an ATX of length $3m$, we show the set S can be partitionned into triplets of sum T. We observe that asking for an explanation of size at most $3m$ is very demanding in this case. Indeed, if $(\mathbf{x}^1, \ldots \mathbf{x}^l)$ supports the claim $(\mathbf{a}, \mathbf{b})$ in a $\mathcal{T}$-ATX, then the sequence $|pros(\mathbf{x}^n, \mathbf{b})|$ decreases by at most one at each step. When applied to a $\mathcal{T}$-ATX supporting the claim $(\mathbf{x}, \mathbf{y})$, denoting $\tau_{j_k \to i_k}^{t_k}$ the transfer performed at step k, this observation entails that exactly one pro $i_k \in [3m]$ is consumed at each step by a transfer whose amount is exactly $t_k = y_{i_k} - x_{i_k} = \widehat{S}_{i_k}$. Moreover, at the end of the sequence, all the surplus $\mathbf{y}_j - \mathbf{x}_j$ for $j \in pros(\mathbf{y}, \mathbf{x}) = [4m] \setminus [3m]$ is consumed. Because $t_k \in S$, $t_k < T/2$ and entirely consuming the surplus associated to agent $j \in [4m] \setminus [3m]$ requires at least 3 transfers, and at most 3 transfers because there are $3m$ transfers and m agents. Thus, an agent $j \in [4m] \setminus [3m]$ participates in exactly 3 transfers, always as a donor, for a total amount of T. Grouping together the values of $\widehat{S}$ that are transferred from the same agent $j \in [4m] \setminus [3m]$ yields a 3-Partition.

Thus, deciding if there is an ATX of size at most ℓ for a given comparative statement in $\mathcal{L}^*$ is at least as hard as deciding if there is a 3-Partition of a given set, i.e. is NP-hard.

A.3 Mixed Integer formalization

We consider the following problem:
Input: two alternatives $\mathbf{a}, \mathbf{b} \in \mathbb{X}^n$, a set of self-evident statements $\mathcal{A} \subseteq \widehat{\mathbb{X}^n}^2$, a positive integer ℓ.
Output: a ℓ-tuple of alternatives $(\mathbf{x}^1, \ldots, \mathbf{x}^{\ell}) \in (\widehat{\mathbb{X}^n})$ such that $((\mathbf{x}^1, \mathbf{x}^2), \ldots, (\mathbf{x}^k, \mathbf{x}^{k+1}), \ldots, (\mathbf{x}^{\ell}, \mathbf{y}))$ is a $\mathcal{A}$-ATX$^{\ell}$ supporting the claim $(\mathbf{a}, \mathbf{b})$ if there is one, or None, else.

As a consequence of Theorem 2, unless $P = NP$, we know that, as soon as $\mathcal{T} \subset \mathcal{A}$ greedy heuristics, or even any tractable algorithm, are bound to yield explanations of suboptimal length. In the light of the results obtained in the paper, we restrict $\mathcal{A}$ to be either $\mathcal{T}$, $\mathcal{T} \cup \mathcal{G}$ or $\mathcal{T} \cup \mathcal{G} \cup cl_{CC}\mathcal{I}$. In order to obtain optimally short explanations,

we reduce the problem of finding a $cl_{CC}(\mathcal{A})$-ATX to a continuous *planning* problem, where the state space is $\widehat{\mathbb{X}^n}$, the initial, current and goal states are respectively $\widehat{\mathbf{a}}$, $\mathbf{x}^k$ and $\widehat{\mathbf{b}}$, and $cl_{CC}(\mathcal{A})$ is the set of actions. This formalism allows to capture:

- $\mathcal{T}$-ATX$^\ell$ with guaranteed failure when $(\mathbf{a},\mathbf{b}) \notin \mathcal{L}^*$ and guaranteed success when $(\mathbf{a},\mathbf{b}) \in \mathcal{L}^*$ and $\ell \geq n$;
- $(\mathcal{T}\cup\mathcal{G})$-ATX$^\ell$ with guaranteed failure when $(\mathbf{a},\mathbf{b}) \notin \mathcal{L}$ and guaranteed success when $(\mathbf{a},\mathbf{b}) \in \mathcal{L}$ and $\ell \geq n+1$;
- $(\mathcal{T} \cup \mathcal{G} \cup cl_{CC}(\mathcal{I}))$-ATX with guaranteed failure when $(\mathbf{a},\mathbf{b}) \notin \mathcal{O}\text{WA}_{\Omega^\mathcal{I}}$ and no guarantee of success.

In turn, we reduce this continuous planning problem to a mixed integer linear optimization (MILO) problem, taking advantage of the powerful dedicated solvers.

We characterize the trade-offs corresponding respectively progressive transfers, gifts or the closure of $\mathcal{I}$ under the rule CC with two linear comparisons:

$$\lambda\mathbf{v}_{low} \leq \mathbf{y}-\mathbf{x} \leq \lambda\mathbf{v}_{up} \tag{1}$$

Lemma 6. • $(x,y) \in \tau^t_{j\to i}$ *iff it satisfies Eq.* (1) *with* $\mathbf{v}_{low} = \mathbf{v}_{up} = \mathbf{e}_i - \mathbf{e}_j$ *and* $\lambda = t$
- $(x,y) \in \mathcal{G}$ *iff it satisfies Eq.* (1) *with* $\mathbf{v}_{low} = \mathbf{0}$, $\mathbf{v}_{up} = \mathbf{1}$ *and any* $\lambda > 0$
- $(x,y) \in cl_{CC}((\mathbf{a}^*,\mathbf{b}^*))$ *iff it satisfies Eq.* (1) *with* $\mathbf{v}_{low} = \mathbf{v}_{up} = \mathbf{b}^* - \mathbf{a}^*$ *and any* $\lambda > 0$

Thus, we represent the set $\mathcal{A}$ by two tuples of vectors $(\mathbf{v}^1_{low},\dots,\mathbf{v}^A_{low})$ and $(\mathbf{v}^1_{up},\dots,\mathbf{v}^A_{up})$ containing respectively the lower and upper bounds characterizing the trade-offs associated to the self-evident comparative statements. If $\mathcal{A} = \mathcal{T}$, $A = \frac{n(n-1)}{2}$. If $\mathcal{A} = \mathcal{T}\cup\mathcal{G}$, $A = \frac{n(n-1)}{2}+1$. If $\mathcal{A} = \mathcal{T}\cup\mathcal{G}\cup cl_{CC}(\mathcal{I})$, $A = \frac{n(n-1)}{2}+1+|\mathcal{I}|$. Our MILO formulation is described by its variables, objective function and linear constraints.

Variables. We use $n\ell$ continuous variables $\mathbf{x}^k_i$, $k\in[\ell], i\in[n]$ for the states, ℓA boolean variables γ^k_m, $k\in[\ell], m\in[A]$ signifying action m is performed as step k, and ℓA continuous variables $\mathbf{c}^k_m$, $k\in[\ell], m\in[A]$ describing the multiplier associated to action m at step k.

Objective. We use a feasibility formulation – it can be described as the minimization of the null function[19].

Constraints. A first set of constraints ensures we remain in the state space $\widehat{\mathbb{X}^n}$, and that we go from the initial state to the goal state.

$$\mathbf{x}^0_i = \widehat{\mathbf{a}}_i \qquad \forall i\in[n] \tag{2a}$$
$$\mathbf{x}^\ell_i = \widehat{\mathbf{b}}_i \qquad \forall i\in[n] \tag{2b}$$
$$\mathbf{x}^k_i \leq \mathbf{x}^k_{i+1} \qquad \forall i\in[n-1]\,\forall k\in[\ell-1] \tag{2c}$$

Equations ensuring exactly one action is performed at each step:

$$\sum_{m\in[A]}\gamma^k_m = 1 \qquad \forall k\in[\ell] \tag{3a}$$
$$0 \leq c^k_m \leq M\gamma^k_m \qquad \forall k\in[\ell],\forall m\in[A] \tag{3b}$$

Equations ensuring the next step is obtained by applying the effects of the chosen action to the current step, by leveraging the previous lemma.

[19] We tried formulating the problem of finding the shortest possible explanation of length upper bounded by ℓ, but the corresponding minimization formulation was less efficient than performing an outer loop over ℓ and solving the feasibility problem.

$$x^{k+1} - x^k \leq \sum_{m\in[A]} c^k_m(\mathbf{v}^m_{up}) \qquad \forall k\in[\ell] \tag{4a}$$
$$x^{k+1} - x^k \geq \sum_{m\in[A]} c^k_m(\mathbf{v}^m_{low}) \qquad \forall k\in[\ell] \tag{4b}$$

A.4 Experimental data generation

We detail the generation process used for the experiments in Table 1. For a given number of agents n, a population of 10 alternative is sampled from $\mathbb{X}^n := [1000]^n$ s.t. the total income of each alternative is equal to $200n$. The results presented are averaged over 10 independent repetitions. To ensure total sum to be exactly equal to $200n$, we used an iterative process for drawing each agent. Suppose we drew the value for $k<n$ agents, whose sum of values is t. The remaining values to be shared between the other $n-k$ agents is then $200n-t$. Let's now compute the maximum and minimum values agent $k+1$ is allowed to take to respect this constraint. First, imagine the agent $k+2$ to n are drawn to their minimum value 0, agent $k+1$ must receive $200n-t$, while remaining in $[1000]$, hence its maximum value is $U_{k+1} = \min(1000, 200n-t)$. Then, if we imagine that the agents $k+2$ to n are drawn to their maximum value 1000, agent $k+1$ must receive the remaining $200n-t-1000*(n-k+1)$. It must also belong to $[1000]$, hence its minimal value is $l_{k+1} = \max(0, 200n-t-1000*(n-k+1))$. Now that we bounded the values for agent $k+1$, we draw a value from the uniform distribution over $\{l_{k+1},\dots,U_{k+1}\}$. This drawn value will update the total attributed sum for agent $k+2$ and we continue until agent n. Once every agent is assigned a value, we order them increasingly. This computation of the bounds of the uniform distribution allows to progressively constrain the values of the remaining agents and handle situations where the first drawn agents are small compared to $200n$, requiring the remaining to receive a high value, and the converse.

A.5 Proof of Theorem 3

The proof is similar to the one provided for Theorem 1.

- $\mathcal{P}_{(t,s,r,m)} \subseteq \mathcal{L}$, because $\mathcal{L}$ satisfies (t), (s) and (r) for the same reasons $\mathcal{L}^*$ does, and also obviously satisfies (m), thus belongs to $\mathfrak{P}_{(t,s,r,m)}$ and refines $\mathcal{P}_{(t,s,r,m)}$.
- $cl_T(\mathcal{S}\cup\mathcal{T}\cup\mathcal{G}) \subseteq \mathcal{P}_{(t,s,r,m)}$ because every relation in $\mathfrak{P}_{(t,s,r,m)}$ contains the set of basic truths $\mathcal{S}\cup\mathcal{T}\cup\mathcal{G}$ and its transitive closure.
- $\bigcup_{\ell=1}^{+\infty}\mathcal{E}(\mathcal{T}\cup\mathcal{G}\text{-ATX}^\ell) \subseteq cl_T(\mathcal{S}\cup\mathcal{T}\cup\mathcal{G})$ because a $(\mathcal{T}\cup\mathcal{G})$ -ATX can structurally be read as a proof by transitivity where the first and last derivations are ordering statements, and all the other are either redistributive transfers or gifts.
- $\mathcal{L} \subseteq \bigcup_{\ell=1}^{n+1}\mathcal{E}(\mathcal{T}\cup\mathcal{G}\text{-ATX}^\ell)$ by Lemma 2.

A.6 Proof of Theorem 4

We recall the counter-example proofing Theorem 4 from the main paper.

Let $\Gamma := \{(z_1,z_2,z_3)\in\mathbb{R}^3$ satisfying (1) $3z_1+2z_2+z_3 \geq 3$ or (2) $z_1\geq 0$ and $z_1+z_2\geq 0$ and $z_1+z_2+z_3\geq 0\}$, and $\mathcal{R}$ the binary relation over $\mathbb{R}^3$ defined by $(\mathbf{x},\mathbf{y})\in\mathcal{R}$ iff $\widehat{\mathbf{y}}-\widehat{\mathbf{x}}\in\Gamma$. $\mathcal{R}$ satisfies (i) and (s) by construction. It is continuous because the set Γ of its acceptable trade-offs is closed (as the union of intersections of preimages of closed sets by continuous functions). It is transitive because Γ is stable under addition (the sum of two vectors satisfying (1) or (2) respectively satisfies (1) or (2), and the sum of one satisfying (1) and

the other (2) satisfies (2)). (r) and (m) are enforced with condition (2). Nevertheless, while the trade-off $t_1 := (2, -4, 6)$ is acceptable (corresponding e.g. to the comparative statement $(0, 8, 10)$ vs $(2, 4, 16)$), the trade off $\frac{1}{2}t_1 = (1, -2, 3)$ is not (while it corresponds e.g. to the comparative statement $(3, 8, 9)$ vs $(4, 6, 12)$).

A.7 Proof of Theorem 5

The proof is similar to those provided for Theorems 1 and 3. It begins with the following soundness results.

- $\mathcal{P}_{(t,r,m,cc,pi^{\mathcal{I}})} \subseteq \mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$ because, for any consistent $\mathcal{I}$, $\Omega^{\mathcal{I}} \subseteq \Omega^{\emptyset}$. The properties (t), (s), (r), (m) and (cc) are satisfied by every precise OWA $\mathcal{O}_{\text{WA}_{\{\omega\}}}$ when $\omega \in \Omega^{\emptyset}$, hence when $\omega \in \Omega^{\mathcal{I}}$. By construction, every precise OWA $\mathcal{O}_{\text{WA}_{\{\omega\}}}$ satisfies (pi$^{\mathcal{I}}$) when $\omega \in \Omega^{\mathcal{I}}$. As those properties are all stable under intersection, $\mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$ satisfies them all.
- $cl_{T,CC}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I}) \subseteq \mathcal{P}_{(t,r,m,cc,pi^{\mathcal{I}})}$ because the basic truths $\mathcal{T}, \mathcal{G}, \mathcal{I}$ belong to every relation satisfying (r), (m) and (pi$^{\mathcal{I}}$), and their deductive closure under rules T and CC is contained by every relation satisfying also (t) and (cc).

To complete the proof, one needs the notion of a CTX (Def. 10) and two additional results:

- soundness of the CTX explanations w.r.t. the formal system $cl_{T,CC}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I})$. Indeed, a CTX explanation is a transitive proof supporting $(\mathbf{x}', \mathbf{y}')$ with arguments that are either basic truths (if they are in $\mathcal{T}$ or $\mathcal{G}$), or deduced from basic truths in $\mathcal{I}$ by the CC rule (if they are in $cl_{CC}(\mathcal{I})$). Then, (x, y) can be derived from $(\mathbf{x}', \mathbf{y}')$ and $(\mathbf{x}, \mathbf{y}) \equiv (\mathbf{x}', \mathbf{y}')$ with the CC rule.
- explainability according to the CTX scheme of every comparative statement in $\mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$, established by Theorem 6 below.

Note that we chose to isolate the characterisation/representation result expressed by Theorem 5 from the notion of explanation, even though explanations are instrumental to its proof. We believe this strong result has a broad scientific interest, in e.g. Economics, independently from the issue of explaining the recommendations of an algorithmic system.

A.8 Proof of Theorem 6

We recall the same constructive proof from the main paper, but in including the case where the alternatives belong to some closed space $[l, U]^n$, with $U \geq l$, not only $\mathbb{R}^n$.

Let $(\mathbf{x}, \mathbf{y}) \in \mathcal{O}_{\text{WA}_{\Omega^{\mathcal{I}}}}$. By Farkas'lemma applied to the MILO formulation of Lemma 5, there are non-negative coefficients $\langle \lambda^*_{(\mathbf{a},\mathbf{b})} \rangle_{(\mathbf{a},\mathbf{b}) \in \mathcal{I}}$, $\langle \mu^*_i \rangle_{i \in [n]}$ and $\langle \nu^*_{i,j} \rangle_{1 \leq i < j \leq n}$ s.t.

$$\widehat{\mathbf{y}} - \widehat{\mathbf{x}} = \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{I}} \lambda_{(\mathbf{a},\mathbf{b})} (\widehat{\mathbf{b}} - \widehat{\mathbf{a}}) + \sum_{i \in [n]} \mu_i \mathbf{e}_i + \sum_{i<j} \nu_{i,j} (\mathbf{e}_i - \mathbf{e}_j)$$

This equation additively decomposes the trade-off corresponding to the comparative statement into 3 terms, each one a summation of trade-offs corresponding to self-evident statements respectively in $cl_{CC}(\mathcal{I})$, $\mathcal{G}$ and $\mathcal{T}$. For every agent i, we split this equation into two parts, one containing the sum of positive values, noted Δ_i^+, and the other containing the sum of negative values, noted Δ_i^-. We then have :

$$\Delta_i^+ = \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{I}} \max\left(0, \lambda^*_j * (\widehat{\mathbf{b}} - \widehat{\mathbf{a}})\right) + \mu^*_i + \sum_{j>i} \nu^*_{i,j}$$
$$\Delta_i^- = \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{I}} \min\left(0, \lambda^*_j * (\widehat{\mathbf{b}} - \widehat{\mathbf{a}})\right) - \sum_{l<i} \nu^*_{l,i}$$

It follows that $\widehat{\mathbf{y}}_i - \widehat{\mathbf{x}}_i = \Delta_i^+ + \Delta_i^-$. Let us first consider the case where $\mathbb{X}^n = \mathbb{R}^n$.

If we define $\widehat{\mathbf{x}}'$ and $\widehat{\mathbf{y}}'$ such that $\forall i \in [n]$:

$$\widehat{\mathbf{x}}'_i = \widehat{\mathbf{x}}_i + \sum_{j=1}^{i} \Delta_{j-1}^+ - \Delta_j^-$$
$$\widehat{\mathbf{y}}'_i = \widehat{\mathbf{y}}_i + \sum_{j=1}^{i} \Delta_{j-1}^+ - \Delta_j^-$$

with $\Delta_0^+ = 0$. One can check that first $\widehat{\mathbf{y}}_i - \widehat{\mathbf{x}}_i = \widehat{\mathbf{y}}'_i - \widehat{\mathbf{x}}'_i$, hence $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$ are congruent. Secondly, one can check that $\forall i \in [n]\ \widehat{\mathbf{x}}'_{i-1} - \widehat{\mathbf{x}}'_i \geq \Delta_{i-1}^+ - \Delta_i^-$, hence the separation between criteria allows $\widehat{\mathbf{x}}'_{i-1}$ to be increased and $\widehat{\mathbf{x}}'_i$ to be decreased by the values given by the Farkas certificate while keeping $\widehat{\mathbf{x}}'_{i-1} \leq \widehat{\mathbf{x}}'_i$. In other words, it allows to perform in any order the $\mathcal{T}$, $\mathcal{G}$ and $cl_{CC}(\mathcal{I})$ statements described by $\langle \lambda^*_{(\mathbf{a},\mathbf{b})} \rangle_{(\mathbf{a},\mathbf{b}) \in \mathcal{I}}$, $\langle \mu^*_i \rangle_{i \in [n]}$ and $\langle \nu^*_{i,j} \rangle_{1 \leq i < j \leq n}$ to build the transitive chain between $\mathbf{x}'$ and $\mathbf{y}'$.

We now consider the case where $\mathbb{X}^n = [l, U]^n$ with $U \geq l$.
If we apply the same construction as the previous case, it can happen that :

$$\max(\widehat{\mathbf{y}}'_n, \widehat{\mathbf{x}}'_n) > U$$

In that case, it is possible to change the definition of $\widehat{\mathbf{x}}'$ and $\widehat{\mathbf{y}}'$ in order to stay in the bounds. Let first compute $c \in]0, 1[$ such that

$$c * \max(\widehat{\mathbf{y}}'_n, \widehat{\mathbf{x}}'_n) < U - l$$

with $\widehat{\mathbf{x}}'$ and $\widehat{\mathbf{y}}'$ the values computed in the case $\mathbb{X}^n = \mathbb{R}^n$. With such a value for c we can build new values for $\widehat{\mathbf{x}}'$ and $\widehat{\mathbf{y}}'$ as :

$$\widehat{\mathbf{x}}'_i = \widehat{\mathbf{x}}'_{i-1} + c * \left(\sum_{k=1}^{i} \Delta_{k-1}^+ - \Delta_i^- \right)$$
$$\widehat{\mathbf{y}}'_i = \widehat{\mathbf{x}}'_{i-1} + c * \left(\sum_{k=1}^{i} \Delta_{k-1}^+ + \Delta_i^+ \right)$$

with $\Delta_0^+ = 0$ and $\mathbf{x}'_0 = l$. In that case $\widehat{\mathbf{y}}'_i - \widehat{\mathbf{x}}'_i = c * (\widehat{\mathbf{y}}_i - \widehat{\mathbf{x}}_i)$, hence $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$ are still congruent. We can also check that $\forall i \in [n]\ \widehat{\mathbf{x}}'_{i-1} - \widehat{\mathbf{x}}'_i \geq c * (\Delta_{i-1}^+ - \Delta_i^-)$, hence the separation between criteria allows $\widehat{\mathbf{x}}'_{i-1}$ to be increased and $\widehat{\mathbf{x}}'_i$ to be decreased by c times the values given by the Farkas certificate while keeping $\widehat{\mathbf{x}}'_{i-1} \leq \widehat{\mathbf{x}}'_i$. We can then produce the same explanation as on the real space, but scaled by the value c in order to remain in $[l, U]^n$.

It results that we can always provide a $(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I})$-CTX of length at most $|\mathcal{I}| + n$.

A.9 Proof of Theorem 7

To prove the first claim, we suppose $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$ are both in the interior of $\widehat{\mathbb{X}^n}$. In this case, the whole line segment from $\widehat{\mathbf{x}}$ to $\widehat{\mathbf{y}}$ is at a positive distance δ from the frontier $\mathbb{F}$. We build an ATX supporting $(\mathbf{x}, \mathbf{y})$

from the decomposition of $\widehat{\mathbf{y}} - \widehat{\mathbf{x}}$ into self-evident trade-offs yielded by Farkas'lemma, that we write shortly as:

$$\widehat{\mathbf{y}} - \widehat{\mathbf{x}} = \sum_{A \in \mathcal{A}} \alpha_A \mathbf{v}^A \tag{5}$$

where the set $\mathcal{A}$ amalgamate all possible actions (see Appendix Section A.3) with $\mathbf{v}^A$ the vector corresponding to the trade-off associated to action A (i.e. $\mathbf{v}^A = \mathbf{b} - \mathbf{a}$ when A is a PI statement, $\mathbf{v}^A = \mathbf{e}_i - \mathbf{e}_j$ with $i < j$ when A is a progressive transfer, and $\mathbf{v}^A \in (\mathbb{R}^+)^n$ when A is a gift). Consider the sequence of alternatives obtained by starting from $\widehat{\mathbf{x}}$ and applying each trade-off. Eq. (5) ensures this sequence ends up in $\widehat{\mathbf{y}}$, but not that every traversed state belongs to $\widehat{\mathbb{X}^n}$. To build an ATX, we leverage the conical structure of self-evident trade-offs. Let K be a positive integer and consider for $0 \leq k \leq K$ the sequence of alternatives $\mathbf{y}^k := \frac{k}{K}\mathbf{x} + \frac{K-k}{K}\mathbf{y}$, each one belonging to the line segment from $\mathbf{x} = y^0$ to $\mathbf{y} = \mathbf{y}^K$, and the following identity:

$$\widehat{\mathbf{y}} - \widehat{\mathbf{x}} = \sum_{k=1}^{K} \left(\sum_{A \in \mathcal{A}} \frac{\alpha_A}{K} \mathbf{v}^A \right) \tag{6}$$

Let $\Delta := \sum_{i \in \mathcal{A}} \|\alpha_A \mathbf{v}^A\|$, so that any sequence of alternatives obtained by chaining those trade-offs starting from any $\mathbf{a} \in \mathbb{X}^n$ remains in the sphere of radius Δ centered around $\mathbf{a}$. By choosing $K \geq \delta/\Delta$, we ensure that each term $\sum_{A \in \mathcal{A}} \frac{\alpha_A}{K} \mathbf{v}^A$ defines a sequence of alternatives remaining safely inside the interior of $\widehat{\mathbb{X}^n}$ and supporting the claim $(\mathbf{y}^k, \mathbf{y}^{k-1})$ in an ATX. Chaining these ATX yield an ATX for $(\mathbf{x}, \mathbf{y})$.

We derive the second claim as a corollary of the first. If the first case does not apply, consider the family of alternatives $\mathbf{x}^\epsilon = \widehat{\mathbf{x}} + \sum_{k=1}^{n} \epsilon^k \mathbf{e}_k$ and $\mathbf{y}^\epsilon = \widehat{\mathbf{y}} - \sum_{k=1}^{n} \epsilon^k \mathbf{e}_k$. There exists a value of ϵ small enough to ensure $\inf_{\omega \in \Omega^{\mathcal{I}}} \omega \cdot (\mathbf{y} - \mathbf{x}^\epsilon) \geq 0$ (i.e. $(\mathbf{x}^\epsilon, \mathbf{y}) \in \mathcal{O}\text{WA}_{\Omega^{\mathcal{I}}}$) and $\mathbf{x}^\epsilon, \mathbf{y}^\epsilon \in \widehat{\mathbb{X}^n} \setminus \mathbb{F}$. Using the first claim, we obtain an ATX for $(\mathbf{x}^\epsilon, \mathbf{y}^\epsilon)$, that we complete with the gifts $(\widehat{\mathbf{x}}, \mathbf{x}^\epsilon)$ and $(y^\epsilon, \widehat{\mathbf{y}})$ in respectively first and last position to obtain an ATX supporting $(\mathbf{x}, \mathbf{y})$.