Explications et caractérisation de décisions équitables

Hénoïk Willot, Sébastien Destercke, Khaled Belahcene

Explications et caractérisation de décisions équitables

Hénoïk Willot[1] **Khaled Belahcene**[2] **Sébastien Destercke**[1]

[1] Heudiasyc, Université de Technologie de Compiègne, France
[2] MICS, CentraleSupélec, Université Paris-Saclay, France
henoik.willot@hds.utc.fr
khaled.belahcene@centralesupelec.fr
sebastien.destercke@hds.utc.fr

Résumé

Les *Ordered weighted averaging* (*OWA*), aussi connues sous le nom *Generalised Gini Index*, sont régulièrement utilisées pour obtenir des décisions équitables. Cependant, bien qu'elles assurent un certain niveau d'équité dans leurs résultats, deux questions subsistent, pourquoi recommandent-elles une alternative donnée et à quel point cette décision est-elle robuste. Nous apportons des outils pratiques et théorisés pour répondre à ces questions, à l'aide d'un moteur d'explications pour les *OWA* robustes qui consiste en une chaîne transitive d'arguments normalisés et considérés évidents, eux-même basés sur les propriétés normatives du modèle. À travers une étude théorique du moteur, nous montrons qu'il est correct (*sound*) et complet par rapport au modèle, et donnons une borne théorique sur la longueur des explications ainsi qu'un algorithme efficace (bien que minimiser la longueur soit NP-difficile). Nous fournissons aussi des éléments montrant que le moteur fonctionne bien sur des données synthétiques. Ainsi, nous garantissons qu'une explication peut toujours être trouvée, et que le raisonnement produit par le moteur explicatif est valide. De plus, ces explications permettent de questionner les fondements du modèle, donc de permettre sa validation et d'établir sa redevabilité, qui sont des composants clés d'une IA de confiance.

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

Équité, Robustesse et Explicabilité sont trois piliers de l'IA de confiance. Les *Ordered Weighted Averaging* (*OWA*) [39], aussi connues sous le nom *generalized Gini indices* [38], sont communément utilisées pour assurer le pilier de l'équité dans différents contextes, par exemple en économie et en choix social computationnel [2], en optimisation multi-objectifs [8], ou en apprentissage de préférences [17] pour n'en citer que quelques uns. Cependant, nous ne connaissons pas de travaux cherchant à expliquer les *OWA* équitables robustes, les *OWA* à poids décroissants définis à partir d'informations partielles. Cela contraste avec d'autres modèles de complexité similaire comme les sommes pondérées robustes, pour lesquelles des moteurs explicatifs corrects (*sound*), complets et efficaces existent.

Nous répondons à ces problèmes, tout d'abord en caractérisant les *OWA* robustes avec des propriétés normatives, puis en utilisant cette caractérisation pour définir un moteur explicatif dont les explications peut être prouvées correctes, complètes et calculables efficacement. Nous avons plusieurs raisons d'adopter ce point de vue normatif et logique :

— il permet un raisonnement réfutable, de construire un système certifié et redevable, aux tiers préjudiciés de contester une décision sur une base formelle ;
— les *OWA* ne sont pas compatibles avec les outils explicatifs comme les *Shapley values* [28] (à cause de

leur symétrie) ou basées sur les gradients [34] (car elles sont hautement non linéaires) ;

— les explications produites reposent uniquement sur l'information donnée et sur les propriétés normatives, sans hypothèse supplémentaire ni divulgation des paramètres du modèle. Ainsi le processus sera plus résilient aux brèches de données personnelles ou à la manipulation, et minimise le biais inductif.

Concrètement, nous souhaitons expliquer des prédictions du type "$\mathbf{x}$ est moins désirable que $\mathbf{y}$", où $\mathbf{x}, \mathbf{y}$ sont des alternatives, ou des états du monde, et où notre fonction de décision doit satisfaire un nombre de propriétés normatives désirables, en plus des préférences données. De plus, nous basons nos conclusions et explications sur les déductions valides pour chaque modèle possible, consistant avec l'information observée. Une telle inférence sceptique est communément utilisée en logique [18] ou dans la décision multi-objectifs [1] et dans l'incertain [35, 29], assurant robustesse dans le sens fort du terme.

Notre proposition commencera avec les propriétés normatives les plus fondamentales, puis nous allons progressivement augmenter leur complexité jusqu'à définir les *OWA* robustes (qui, à leur limite, incluent les *OWA* précis). Pendant ces étapes, nos résultats vont montrer pourquoi expliquer les *OWA* équitables robustes a reçu peu voir aucune attention : le problème est loin d'être trivial, du point de vue théorique ou algorithmique, et n'est pas une simple adaptation de résultats existants, comme ceux de la somme pondérée. Notre point de départ est le problème de la comparaison d'alternatives, décrites par un nombre de points de vue exprimés sur une échelle commensurable à l'aide d'une structure de préférences qui vérifie des propriétés fortes de la théorie de la décision (transitivité, symétrie, redistributivité et monotonie) qui sont normativement désirables pour le procédé de décision étudié. De telles structures de préférences sont fortement liées à la notion de dominance de Lorenz généralisée, introduite il y a longtemps dans le domaine de l'économie du bien-être [36].

Dans la section 4, nous nous intéressons aux OWA équitables robustes (c.-à-d. avec des poids décroissants), qui raffinent les préférences de la dominance de Lorenz généralisée, résolvant le problème qu'a cette dernière à régulièrement ne pas pouvoir comparer les alternatives, c.-à-d. être trop indécisive. Pour ce faire, nous considérons qu'en plus de devoir satisfaire les propriétés théoriques précédemment mentionnées, un utilisateur a émis des préférences, sous forme de paire comparative obtenues par exemple avec de l'apprentissage actif [7], mais qui ne permettent d'identifier qu'un sous-ensemble de modèles possibles.

Nos principales contributions sont les suivantes :

— A partir de la littérature sur les explications de la dominance de Lorenz, nous montrons que trouver l'explication optimale (la plus courte) sous forme d'un ensemble successif de transferts est un problème NP-difficile et nous donnons une heuristique ayant de meilleures performances empiriques que les précédentes.

— Nous proposons, ce qui a notre connaissance est, une nouvelle axiomatique pour des ensembles convexes d'OWA (que nous nommons OWA équitables robustes), qui inclue la dominance de Lorenz et l'index de Gini généralisé dans un cadre unique. De plus, ces axiomes étant plutôt naturels nous permettent de définir des méchanismes explicatifs. Cela contraste avec les axiomatiques qui ont besoin d'axiomes techniques comme la continuité qui sont difficiles à utiliser dans une explication. Nous montrons aussi que nos explications sont logiquement correctes et complètes, c.-à-d. que toutes les préférences peuvent être expliquées et que toutes les préférences expliquées sont vraies. Nous proposons aussi des heuristiques pour fournir rapidement des explications.

2 Préliminaires

Nous étudions les préférences entre *alternatives*, décrite par plusieurs attributs mesurés sur une échelle commune : nous noterons $[n] = \{1, \ldots, n\}$ l'ensemble des attributs et par $\mathbb{X}$ cette échelle commune, étant un intervalle non trivial des réels. Les alternatives $\mathbf{x} \in \mathbb{X}^n$ sont décrites par une minuscule. $[n]$ peut représenter des points de vues dans différents contextes, comme la décision multi-critère ou multi-agent, et les alternatives sont des choix possibles décrits par ces points de vues, par exemple la distribution de richesse dans un groupe d'agents. Nous noterons $(\mathbf{e}^1, \ldots, \mathbf{e}^n)$ pour la base canonique de $\mathbb{R}^n$ et $\widehat{\mathbb{X}^n}$ le sous-ensemble de $\mathbb{X}^n$ des n-uplets triés par ordre croissant.

Les préférences sont représentées par une relation binaire $\mathcal{R}$ sur $\mathbb{X}^n$. Étant donné deux alternatives $\mathbf{x}, \mathbf{y}$ de $\mathbb{X}^n$, la *paire comparative* $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ (ou $\mathbf{x}\ \mathcal{R}\ \mathbf{y}$) dénote que l'alternative $\mathbf{x}$ est au plus aussi désirable que l'alternative $\mathbf{y}$. Par conséquence, il existe quatre possibilités quand nous comparons $\mathbf{x}$ à $\mathbf{y}$: *(i)* $\mathbf{y}$ est strictement préféré à $\mathbf{x}$ si $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ et $(\mathbf{y}, \mathbf{x}) \notin \mathcal{R}$; *(ii)* $\mathbf{x}$ est strictement préféré à $\mathbf{y}$ si $(\mathbf{x}, \mathbf{y}) \notin \mathcal{R}$ et $(\mathbf{y}, \mathbf{x}) \in \mathcal{R}$; *(iii)* $\mathbf{x}, \mathbf{y}$ sont indifférents quand $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ et $(\mathbf{y}, \mathbf{x}) \in \mathcal{R}$ [1] ; *(iv)* $\mathbf{x}, \mathbf{y}$ sont *incomparables* quand $(\mathbf{x}, \mathbf{y}), (\mathbf{y}, \mathbf{x}) \notin \mathcal{R}$. Quelques ensembles de préférences nous intéressent tout particulièrement :

Définition 1 (réarrangement $\mathcal{S}$)**.** Soit $\mathcal{S}$ l'ensemble de paires comparatives $(\mathbf{x}, \mathbf{y})$ t.q. $\mathbf{y}$ est une permutation de $\mathbf{x}$. $\mathcal{S}$ est une relation d'équivalence et chaque alternative $\mathbf{x} \in \mathbb{X}^n$ a un unique équivalent $\widehat{\mathbf{x}}$ par $\mathcal{S}$ dans l'ensemble $\widehat{\mathbb{X}^n}$.

Définition 2 (transferts $\mathcal{T}$)**.** Soient $t \in \mathbb{R}$, $i, j \in [n]$ et la paire comparative $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$, où les deux alternatives sont triées dans l'ordre croissant, et où $\widehat{\mathbf{y}}$ est la situation où, en

1. Pour ce faire, $\mathcal{R}$ est évidemment *reflexive*, c.-à-d. $(\mathbf{x}, \mathbf{x}) \in \mathcal{R}$. Nous faisons cette hypothèse tout au long de l'article.

commençant par $\widehat{\mathbf{x}}$, la quantité t est prise de l'agent j et donnée à l'agent i, c.-à-d. $\widehat{\mathbf{y}} = \widehat{\mathbf{x}} + t\mathbf{e}^i - t\mathbf{e}^j$, noté $\tau^t_{j\to i}$. Quand $t > 0$ et $i < j$, ce transfert est dit *redistributif* [2], et notons $\mathcal{T}$ l'ensemble de tous les transferts redistributifs, c.-à-d. $\mathcal{T} := \bigcup_{t>0} \bigcup_{1\le i<j\le n} \tau^t_{j\to i}$.

Définition 3 (dons $\mathcal{G}$)**.** Soit $\mathcal{G} := \{(\mathbf{x}, \mathbf{y}) : \forall i \in [n]\ \mathbf{x}_i \le \mathbf{y}_i\}$, l'ensemble des paires comparatives où les préférences des agents de $[n]$ sont unanimes [3]

Nous donnons un exemple général, où une autorité centrale doit redistribuer des revenus d'investissements, permettant d'illustrer le problème considéré, les notions impliquées et d'appliquer notre solution. Nous invitons le lecteur à revenir vers lui au cours de sa lecture.

Exemple 1 (Exemple illustratif)**.**

Inv.	*Alice*	*Bob*	*Charlie*	*David*	*Emma*
a	*31*	*83*	*70*	*16*	*51*
b	*28*	*98*	*25*	*2*	*84*
c	*22*	*23*	*76*	*82*	*34*
d	*96*	*6*	*18*	*17*	*88*

Nous commençons par affirmer que les alternatives doivent être anonymisées et triées par ordre croissant de satisfaction.

Inv.	*#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*

Supposons que l'autorité centrale a déjà statué sur le fait que **b** *est au plus aussi désirable que* **d** *et qu'elle utilise un OWA précis* [4] $OWA_{\{\omega^*\}}$ *avec* $\omega^* = (.70, .10, .10, .05, .05)$*. Comment peut-on expliquer pourquoi* **c** *est préféré à* **a** *tout en gardant* ω^* *secret ?* [5]

Pour prouver cette préférence, nous construisons une chaîne d'alternatives $(\mathbf{a}, \widehat{\mathbf{a}}, \mathbf{x}^1, \mathbf{x}^2, \widehat{\mathbf{c}}, \mathbf{c})$*, avec* $\mathbf{x}^1 :=$ (16 35 47 70 83) *et* $\mathbf{x}^2 :=$ (22 23 32 76 80)*.* $\mathbf{x}^1$ *doit être considéré meilleur que* $\widehat{\mathbf{a}}$*, car il est obtenu en transférant* 4 *unités du troisième agent le moins satisfait vers le second agent le moins satisfait. La situation* $\mathbf{x}^2$ *doit être considérée meilleure que* $\mathbf{x}^1$*, vu que le changement* $(+6, -12, -15, +6, -3)$ *de* $\mathbf{x}^1$ *vers* $\mathbf{x}^2$ *doit être considéré positif, comme ce dernier correspond* ceteris paribus *à une fois et demie le changement de* $\widehat{\mathbf{b}}$ *vers* $\widehat{\mathbf{d}}$*. Enfin,* $\widehat{\mathbf{c}}$ *doit être considéré meilleur que* $\mathbf{x}^2$*, car chaque agent est au moins aussi satisfait. Ainsi, la transitivité des préférences permet de dire que* **c** *est préféré par rapport à* **a***. Techniquement, les préférences* $(\mathbf{a}, \widehat{\mathbf{a}})$ *et* $(\widehat{\mathbf{c}}, \mathbf{c})$ *sont des réarrangements,* $(\widehat{\mathbf{a}}, \mathbf{x}^1)$ *est un transfert redistributif et* $(\mathbf{x}^2, \widehat{\mathbf{c}})$ *est un don.*

3 Expliquer la dominance de Lorenz

Notre but est de caractériser nos inférences sceptiques et moteurs explicatifs pour des règles de décision équitables, à commencer par la dominance de Lorenz restreinte puis sa version généralisée, définies par :

Définition 4 (Dominances de Lorenz $\mathcal{L}$ et $\mathcal{L}^*$)**.** La *dominance de Lorenz généralisée* est la relation binaire $\mathcal{L}$ définie sur $\mathbb{X}^n$ telle que $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}$ ssi $\forall i \in [n]$, $\sum_{k=1}^{i} \widehat{\mathbf{x}}_k \le \sum_{k=1}^{i} \widehat{\mathbf{y}}_k$. La *dominance de Lorenz restreinte* est le sous-ensemble $\mathcal{L}^*$ de $\mathcal{L}$ où $\mathbf{x}, \mathbf{y}$ ont le même revenu total, c.-à-d. $\sum_{k\in[n]} \mathbf{x}_k = \sum_{k\in[n]} \mathbf{y}_k$.

3.1 Sémantique des préférences sceptiques

Dans cet article, nous nous intéressons aux relations de préférences qui vérifient un certain nombre de propriétés issues de la théorie de la décision.

Propriété (t). $\mathcal{R}$ est *transitive* quand, pour tous $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{X}^n$, si $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ et $(\mathbf{y}, \mathbf{z}) \in \mathcal{R}$, alors $(\mathbf{x}, \mathbf{z}) \in \mathcal{R}$.

Propriété (s). $\mathcal{R}$ est *symétrique* quand elle inclut (ou raffine) $\mathcal{S}$, c.-à-d. $\mathcal{R} \supseteq \mathcal{S}$ contient tous les réarrangements.

Propriété (r). $\mathcal{R}$ est *redistributive* quand elle inclut $\mathcal{T}$, c.-à-d. $\mathcal{R} \supseteq \mathcal{T}$ contient tous les transferts redistributif.

Notons $\mathfrak{P}_{(t,s,r)}$ l'ensemble de toutes les relations binaires réflexives sur $\mathbb{X}^n$ vérifiant simultanément ces trois propriétés. En assumant que cet ensemble est non vide (montré dans la prochaine section), soit $\mathcal{P}_{(t,s,r)}$ la relation de préférence définie comme l'intersection de toutes les relations de $\mathfrak{P}_{(t,s,r)}$. Comme les trois propriétés sont stables par intersection [6], $\mathcal{P}_{(t,s,r)}$ appartient à $\mathfrak{P}_{(t,s,r)}$ et est son plus petit élément du point de vue de l'inclusion. Par conséquent, du point de vue sémantique, $\mathfrak{P}_{(t,s,r)}$ peut être vu comme l'ensemble des mondes possibles et $\mathcal{P}_{(t,s,r)}$ comme l'ensemble des décisions qui peuvent être inférées sceptiquement, c.-à-d. qui se produisent dans tous les mondes possibles. $\mathfrak{P}_{(t,s,r)}$ peut aussi être vu comme un jury, où les jurés sont toutes les relations de préférences qui respectent les propriétés normatives, et $\mathcal{P}_{(t,s,r)}$ est l'ensemble des décisions unanimes en leur sein. Ces décisions prudentes sont nécessaires du

2. À interpréter : "prendre une quantité $t > 0$ de l'agent j et la donner à l'agent plus pauvre i" (en conservant l'ordre social). De tels transferts sont aussi appelés transferts de *Pigou-Dalton* ou de *Robin des Bois*.

3. Il s'agit simplement de la dominance de Pareto de $\mathbf{y}$ sur $\mathbf{x}$, où $\mathbf{y}_i = \mathbf{x}_i + t_i$, $t_i > 0$ étant le don fait à l'agent i

4. La définition de sa représentation numérique est rappelée au début de la section 4.1.

5. Nous pouvons voir que calculer l'importance individuelle des agents, par exemple de Bob avec la dérivée partielle $\partial_{\text{Bob}} OWA_{\{\omega^*\}}(\mathbf{a}) = .70$, attribue une haute importance à Bob ce qui est trompeur, tandis que les valeurs de Shapley des joueurs Alice, Bob, Charlie, David et Emma sont égales.

6. dans le sens où si $\mathcal{R}_1$ et $\mathcal{R}_2$ satisfont simultanément cette propriété, alors $\mathcal{R}_1 \cap \mathcal{R}_2$ aussi.

point de vue des propriétés normatives et ne sont donc pas arbitraires ou contingentes. En se concentrant sur ces décisions, nous offrons de la robustesse à l'utilisateur, et nous espérons les accompagner de preuves et d'explications basées sur ces propriétés normatives.

3.2 Une représentation numérique de $\mathcal{L}^*$

Il est facile de vérifier que $\mathcal{L}$ et $\mathcal{L}^*$, définies dans Def. 4, satisfont les propriétés (t), (s) et (r). $\mathfrak{B}_{(t,s,r)}$ est donc non vide, et, comme nous le verrons, $\mathcal{L}^*$ est son plus petit élément, soit $\mathcal{P}_{(t,s,r)} = \mathcal{L}^*$. C'est un résultat fort, permettant de dire qu'une paire comparative $(\mathbf{x}, \mathbf{y})$ est obtenue par toutes les relations de préférences vérifiant (t), (s) et (r) simplement en triant $\mathbf{x}$ et $\mathbf{y}$ et en calculant leur sommes cumulatives et comparant (au plus) n paires d'éléments.

3.3 Un calcul correct et complet des préférences

Nous nous intéressons à créer un système déductif formel qui reflète ces propriétés normatives, et qui infère des paires comparatives à partir d'autres.

Inférence. Nous associons la propriété (t) à la règle

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

Vérités premières [7]**.** Pour refléter les propriétés (s) et (r), nous considérons les réarrangements $\mathcal{S}$ et les transferts redistributifs $\mathcal{T}$ comme des vérités premières.

Soit $cl_T(\mathcal{S} \cup \mathcal{T})$ la clôture déductive [8] de l'ensemble des vérités premières $\mathcal{S} \cup \mathcal{T}$ par l'opérateur T, c.-à-d. l'ensemble de toutes les paires comparatives qui peuvent être prouvées à partir de prémisses dans $\mathcal{S}$ ou dans $\mathcal{T}$ en enchaînant les déductions à partir de la règle T. La *correction* (*soundness*) du système formel par rapport à la sémantique, c.-à-d. $cl_T(\mathcal{S} \cup \mathcal{T}) \subseteq \mathcal{P}_{(t,s,r)}$, signifiant que toute décision qui peut être prouvée peut aussi être sceptiquement inférée, est obtenue immédiatement par la construction des règles et vérités premières qui reflètent les propriétés des préférences. La *complétude*, c.-à-d. $\mathcal{P}_{(t,s,r)} \subseteq cl_T(\mathcal{S} \cup \mathcal{T})$, signifiant que toute paire qui ne peut être réfutée empiriquement peut être prouvée, provient du fait que $cl_T(\mathcal{S} \cup \mathcal{T})$ satisfait (t) car elle est fermée sous T, et évidemment (s) et (r).

3.4 Explications schématiques

Bien que le résultat de complétude soit satisfaisant, les preuves résultantes seront toujours sous forme d'arbres, qui ne seront sûrement pas assez concis ou simples pour être cognitivement acceptées par des agents. Nous proposons donc des explications sous forme d'une *chaîne transitive d'arguments* formant une séquence "d'actes de langage".

Définition 5 (Moteur ATX)**.** Soit $\mathcal{R}$ une relation binaire sur $\widehat{\mathbb{X}^n}$. Une *explication anonyme et transitive basée sur les vérités dans* $\mathcal{R}$ de longueur ℓ ($\mathcal{R}$-ATX$^\ell$) est une paire (s, c) où le *support* s est un ℓ-uplet de paires comparatives $s = ((\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^\ell, \mathbf{y}^\ell)) \in \mathcal{R}^\ell$ et *l'affirmation* c est une paire comparative $c = (\mathbf{x}, \mathbf{y})$ vérifiant $\mathbf{x}^1 = \widehat{\mathbf{x}}$, $\mathbf{y}^\ell = \widehat{\mathbf{y}}$ et, pour tout entier k entre 2 et ℓ, $\mathbf{x}^k = \mathbf{y}^{k-1}$.

Soit $\ell \in \mathbb{N}^*$ et soit $\mathcal{E}(\mathcal{T}\text{-ATX}^\ell)$ l'ensemble de toutes les *décisions explicables*, c.-à-d. des affirmations associées à une chaîne de vérités premières de $\mathcal{T}$ par une ATX de taille ℓ. Cet ensemble est inclus dans $cl_T(\mathcal{S} \cup \mathcal{T})$, car une ATX justifiant la conclusion $(\mathbf{x}, \mathbf{y})$ peut être vue comme une chaîne transitive entre $\mathbf{x}$ et $\mathbf{y}$, où la première (c.-à-d. $(\mathbf{x}, \widehat{\mathbf{x}})$) et la dernière (c.-à-d. $(\widehat{\mathbf{y}}, \mathbf{y})$) paire comparative appartiennent à $\mathcal{S}$, et toutes les autres à $\mathcal{T}$: les $\mathcal{T}$-ATX sont correctes par rapport à $cl_T(\mathcal{S} \cup \mathcal{T})$. Sont-elles complètes ? En effet, elles sont même complètes par rapport à $\mathcal{L}^*$, d'après ce résultat bien connu des années 1930.

Lemme 1 (Hardy et al. 1929)**.** $\mathcal{L}^* \subseteq \bigcup_{\ell=1}^n \mathcal{E}(\mathcal{T}\text{-}ATX^\ell)$

Nous rappelons rapidement la preuve constructive, étant donné qu'elle forme un algorithme que nous améliorerons.

Ébauche de preuve. Initialiser $\mathbf{x}^0 := \widehat{\mathbf{x}}$; A l'étape k trouver i^*, j^*, t^* t.q. $j^* = \arg\min_j \mathbf{x}_j^{k-1} > \widehat{\mathbf{y}}_j$, $i^* = \arg\max_{i:i<j} \mathbf{x}_i^{k-1} < \widehat{\mathbf{y}}_i$ et $t^* = \min(\widehat{\mathbf{y}}_{i^*} - \mathbf{x}_{i^*}^{k-1}, \mathbf{x}_{j^*}^{k-1} - \widehat{\mathbf{y}}_{j^*})$ définir $\mathbf{x}^k$ t.q. $(\mathbf{x}^{k-1}, \mathbf{x}^k) \in \tau_{j^* \to i^*}^{t^*}$. Le nombre d'agents $i \in [n]$ t.q. $\mathbf{x}_i^k \neq \widehat{\mathbf{y}}_i$ décroît strictement avec k, d'où la terminaison de l'algorithme qui délivre une $\mathcal{T}$-ATX de taille au plus n pour $(\mathbf{x}, \mathbf{y})$. □

Comme les $\mathcal{T}$-ATX sont correctes et complètes, nous obtenons le résultat suivant

Théorème 1 (Explicabilité de la dominance de Lorenz restreinte)**.**

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

Exemple 2. *Considérons les alternatives* $\widehat{\mathbf{c}}$ *et* $\widehat{\mathbf{b}}$ *de l'exemple 1. Calculer leur sommes cumulées montre que* $(\mathbf{b}, \mathbf{c}) \in \mathcal{L}^*$. *La plus petite explication supportant* $(\mathbf{b}, \mathbf{c})$ *est de longueur 4 :*

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

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

7. Aussi appelées *axiomes*, mais nous évitons cette dénomination qui change de sens selon l'utilisation en logique ou en théorie de la décision.

8. C'est aussi la clôture transitive vu que le seul opérateur est la transitivité, ce qui est contingent aux propriétés de la dominance de Lorenz.

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

FIGURE 1 – Notre algorithme – le choix du donneur j^*, receveur i^* et de la quantité t^* pour le transfert à l'étape k.

3.5 Aspects computationnels

Ainsi, étant donné une paire comparative $(\mathbf{x}, \mathbf{y})$, il est équivalent de *(i)* décider si elle est obtenue pour toute relation de préférence vérifiant (t), (s) et (r) ; *(ii)* chercher pour une preuve déductive utilisant la règle T avec les prémisses issues de $\mathcal{S}$ et $\mathcal{T}$; *(iii)* résoudre le problème de trouver une explication ; ou *(iv)* ordonner - sommer - comparer à l'aide de la représentation numérique de $\mathcal{L}^*$. Évidemment, la dernière est la plus facile d'un point de vue computationnel. En effet, nous prouvons que le problème de trouver une explication d'une taille donnée est difficile.

Théorème 2 (Difficulté de trouver des explications courtes)**.** *Étant donné* $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}^*$ *et un entier positif* k*, décider s'il existe une* $\mathcal{T}$*-ATX de taille au plus* k *pour* $(\mathbf{x}, \mathbf{y})$ *est NP-difficile.*

Ébauche de preuve. Réduction depuis le problème de 3-partition [19]. Soit S un ensemble d'entiers t.q. $|S| = 3m$, $\sum_{s \in S} s = mT$ et $\frac{T}{4} < s < \frac{T}{2}$ $\forall s \in S$, entrée du problème de 3-partition. Nous construisons les alternatives $\mathbf{x}$ et $\mathbf{y} \in \widehat{\mathbb{X}^{4m}}$ t.q. $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}^*$, définies par $\mathbf{x}_i = \sum_{j=1}^{i-1} \widehat{S}_j$ si $1 \leq i \leq 3m$ et $\mathbf{x}_i = iT$ si $3m < i \leq 4m$ et par $\mathbf{y}_i = \sum_{j=1}^{i} \widehat{S}_j$ si $1 \leq i \leq 3m$ et $\mathbf{y}_i = (i-1)T$ si $3m < i \leq 4m$. □

La difficulté de résolution nous amène à considérer :

— Une formulation en programmation linéaire mixte (MILP) sous la forme d'un problème de planification continue [9] dont la solution est l'explication la plus courte.

— Une heuristique gloutonne, semblable à l'algorithme HLP [22] sous-jacent au Lemme 1, mais orienté vers des explications courtes, est décrite par ses choix de valeurs pour i^*, j^*, t^* à l'étape k dans la figure 1.

Les résultats expérimentaux menés sur des données synthétiques sont donnés dans la table 1. Pour un nombre donné n d'agents, un ensemble de 10 alternatives est échantillonné de $\mathbb{X}^n := [1000]^n$ t.q. la richesse totale de chaque alternative soit égale à $200n$. Les résultats données, obtenus sur un ordinateur portable standard, sont moyennés sur 10 répétitions indépendantes. La table 1 montre que notre heuristique est très rapide et plutôt proche de l'optimum. Les résultats suggèrent que la taille optimale croît selon $0.6n$, tandis que notre heuristique croît selon $0.7n$.

9. *L'espace d'états* est $\widehat{\mathbb{X}^n}$, les états *initial, objectif* et *actuels* sont respectivement $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$ et $\mathbf{x}^k$, et les *actions* sont les vérités premières.

3.6 Le cas de la dominance de Lorenz généralisée

La machinerie que nous avons construit ne permet pas de comparer des populations qui ont une richesse totale différente. Pour cette raison, les économistes ont proposé de considérer les *dons* comme désirables.

Propriété (m). $\mathcal{R}$ est *monotone* quand $\mathcal{R} \supseteq \mathcal{G}$.

Cette propriété est une sorte de garantie *d'efficacité* : dépenser le surplus est préférable à ne pas le dépenser [10]

Sémantique et représentation. Nous pouvons observer que $\mathcal{L}$ vérifie (m) (alors que $\mathcal{L}^*$ non), ainsi l'ensemble $\mathfrak{P}_{(t,s,r,m)}$ contenant toutes les relations binaires réflexives satisfaisant conjointement (t), (s), (r) et (m) est non vide. Soit $\mathcal{P}_{(t,s,r,m)}$ l'intersection de toutes les relations de $\mathfrak{P}_{(t,s,r,m)}$, qui est le plus petit élément de $\mathfrak{P}_{(t,s,r,m)}$, comme la monotonie est aussi stable par intersection. $\mathfrak{P}_{(t,s,r,m)}$ est un sous-ensemble (strict) de $\mathfrak{P}_{(t,s,r)}$, ainsi $\mathcal{P}_{(t,s,r,m)}$ raffine $\mathcal{P}_{(t,s,r)}$: ajouter une nouvelle propriété réduit le spectre de mondes possibles et réduit la possibilité de trouver un contre-argument à une préférence, permettant donc d'augmenter le nombre de décision sceptiques.

Déduction. Prendre en compte (m) dans le système déductif est simple : il suffit de considérer les dons $\mathcal{G}$ comme vérités premières, en plus des réarrangements $\mathcal{S}$ et des transferts $\mathcal{T}$. La clôture déductive par T est notée $cl_T(\mathcal{S} \cup \mathcal{T} \cup \mathcal{G})$.

Explications. Nous gardons la structure des explications anonymes et transitives, et augmentons leur pouvoir expressif en ajoutant les éléments de l'ensemble $\mathcal{G}$ des dons en plus des transferts redistributifs de l'ensemble $\mathcal{T}$.

Résultats structurels. Comme la dominance de Lorenz généralisée $\mathcal{L}$ appartient à $\mathfrak{P}_{(t,s,r,m)}$, elle raffine $\mathcal{P}_{(t,s,r,m)}$. $cl_T(\mathcal{S} \cup \mathcal{T} \cup \mathcal{G})$ est clairement correct et complet par rapport à $\mathcal{P}_{(t,s,r,m)}$. Les $(\mathcal{T} \cup \mathcal{G})$-ATX sont correctes *by design* vis à vis de $cl_T(\mathcal{S} \cup \mathcal{T} \cup \mathcal{G})$. Cette inclusion des relations de préférences se réduit grâce au résultat suivant.

Lemme 2 (Chong, 1976)**.** $\mathcal{L} \subseteq \bigcup_{\ell=1}^{n+1} \mathcal{E}(\mathcal{T} \cup \mathcal{G}\text{-}ATX^\ell)$

Ébauche de preuve. Il suffit d'appliquer un don en première ou en dernière position, de manière à avoir $\mathbf{x}^1$ et $\mathbf{y}$ avec le même revenu total, et ensuite de chercher une séquence de transferts redistributifs de $\mathbf{x}^1$ vers $\mathbf{y}$ via le Lemme 1. □

Théorème 3 (Explicabilité de la dominance de Lorenz généralisée)**.**

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

10. Dans certains contextes, cette notion est connue pour être aux antipodes de l'équité [12].

	Long. Moy.			% d'égal./vict./déf. entre les longueurs des méthodes					Temps Moy. (s)			Time-out
n	◊	□	△*	◊ = □	◊ < □	◊ > □	□ = △*	□ > △*	◊	□	△	△ %
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

◊ Algorithme HLP □ Notre algorithme △ Optimum

* la valeur de l'heuristique est choisie comme référence si le temps de calcul de l'optimum a dépassé le timeout (150s par explication).

TABLE 1 – Comparaison entre notre heuristique et l'algorithme HLP pour $\mathcal{L}^*$

Aspects computationnels. Du point de vue théorique, le problème de trouver des explications courtes pour la dominance de Lorenz généralisée est au moins aussi dur que pour la dominance restreinte. Curieusement, du point de vue algorithmique, l'heuristique découlant du Lemme 2, c.-à-d. systématiquement se réduire au problème avec un revenu total égal par un don initial (ou final) est sous-optimal, comme le montre l'exemple suivant.

Exemple 3. *Prenons les alternatives* $\widehat{\mathbf{d}}$ *et* $\widehat{\mathbf{c}}$ *de l'exemple 1. Calculer leurs cumulées montre que* $(\mathbf{d}, \mathbf{c}) \in \mathcal{L}$*. L'explication la plus courte supportant* $(\mathbf{d}, \mathbf{c})$ *est de taille 3 :*

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

Elle est strictement plus courte que toutes les explications plaçant le don en première ou dernière position.

4 Expliquer les décisions des *OWA* équitables et robustes

Bien qu'étant des relations de préférences équitables fondamentales, les dominances de Lorenz restent très indécises. Dans une situation de prise de décision [13] nous avons besoin d'une structure de préférences plus résolue, par exemple pour faire un choix (sélectionner une alternative favorite) ou donner un classement des alternatives.

De plus, la dominance de Lorenz est par définition non paramétrée, mais il peut être utile de considérer des raffinement paramétrés, capturant des motifs de préférences plus spécifiques tout en permettant des explications simples. Ensuite, nous allons complémenter les principes normatives avec *de l'information préférentielle*, permettant à la fois de restreindre l'ensemble des mondes possibles et d'augmenter la base du raisonnement et des explications.

4.1 Préférences basées sur un ensemble d'*OWA*

Les préférences représentées par la fonction de score nommée *Ordered weighted averaging* proviennent de [38] dans le contexte des indices d'inégalité et de [39] dans le contexte de la décision multi-critères. L'*OWA* est paramétrisée par un n-uplet $\boldsymbol{\omega}$ et associe à l'alternative $\mathbf{x}$ au score $\sum_{i\in[n]} \boldsymbol{\omega}_i \widehat{\mathbf{x}}_i$, qui est une somme pondérée ordonnée.

Le même modèle est parfois appelé *generalized Gini index* (*GGI*), comme une valeur spécifique du paramètre $\boldsymbol{\omega}$ produit l'indice de Gini. Nous donnons une définition basée sur la structure de préférence plutôt que sur le score, qui représente intrinsèquement l'inférence sceptique sur un ensemble Ω de valeurs du paramètre qui représente l'information préférentielle incomplète.

Définition 6 (Préférences basées sur les *OWA* robustes)**.** Soit Ω un sous-ensemble non vide de la sphère unité L_1 [11] de $\mathbb{R}^n$ et soit O_{WA_Ω} la relation binaire définie par $(\mathbf{x}, \mathbf{y}) \in O_{\mathrm{WA}_\Omega}$ ssi $\sum_{i\in[n]} \boldsymbol{\omega}_i \widehat{\mathbf{x}}_i \leq \sum_{i\in[n]} \boldsymbol{\omega}_i \widehat{\mathbf{y}}_i$ pour tout $\boldsymbol{\omega} \in \Omega$.

4.2 Propriétés des préférences basées sur les *OWA*

Nous observons que $O_{\mathrm{WA}_{\{\boldsymbol{\omega}\}}}$ est :
- réflexive, transitive et symétrique qu'importe $\boldsymbol{\omega}$;
- monotone quand toutes les valeurs de $\boldsymbol{\omega}$ sont non négative, reflétant l'attrait de tous les critères ;
- redistributive quand les valeurs de $\boldsymbol{\omega}$ sont décroissantes : les agents les moins satisfaits sont plus importants ;

Soit [12] $\Omega^{\emptyset}$ l'ensemble de tous les n-uplets non-négatifs, décroissants, non-nuls de la sphère unité de $\mathbb{R}^n$.

Lemme 3 (Argyris et al. 2022)**.** $\mathcal{L} = O_{WA_{\Omega^{\emptyset}}}$

De plus, plusieurs caractérisations des préférences basées sur des *OWA* ont été proposées (par exemple [38, 6, 32]). Elles diffèrent légèrement, mais elles requièrent toutes que la relation soit décisive et continue d'une certaine façon pour assurer qu'elle soit représentée par une fonction de valeur réelles, et impose une condition pour assurer que cette fonction est additive sur l'ensemble $\widehat{\mathbb{X}^n}$. Nous détaillons les résultats obtenus par Ben-Porath et Gilboa.

Propriété (d). $\mathcal{R}$ est *décisive* quand pour toutes alternatives $\mathbf{x}, \mathbf{y}$, $(\mathbf{x}, \mathbf{y})$ ou $(\mathbf{y}, \mathbf{x})$ (ou les deux, où dans ce cas $\mathbf{x}$ et $\mathbf{y}$ sont considérées aussi désirable l'un que l'autre) appartiennent à $\mathcal{R}$.

11. Cette condition assure la non trivialité des préférences (vu que le paramètre nul correspond à l'indifférence totale) et la non redondance des paramètres (comme une transformation linéaire des paramètres induit la même relation), tandis que le choix de la norme L_1 assure la calculabilité.

12. Cette notation est consistante avec la Def. 8.

O_{WA_Ω} est décisive ssi Ω est un singleton.

Propriété (c). $\mathcal{R}$ est *continue* quand, pour toute alternative $\mathbf{z}$, les ensembles $\{\mathbf{x} : (\mathbf{x}, \mathbf{z}) \in \mathcal{R}\}$ et $\{\mathbf{y} : (\mathbf{z}, \mathbf{y}) \in \mathcal{R}\}$ sont fermés.

Propriété (i). $\mathcal{R}$ est *invariante* (par rapport aux dons préservant l'ordre social) quand, pour toutes alternatives $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$, s'il existe un agent $i \in [n]$ et $t \in \mathbb{R}$ t.q. $\widehat{\mathbf{x}}' = \widehat{\mathbf{x}} + t\mathbf{e}^i$ et $\widehat{\mathbf{y}}' = \widehat{\mathbf{y}} + t\mathbf{e}^i$, alors $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) \in \mathcal{R} \iff (\widehat{\mathbf{x}}', \widehat{\mathbf{y}}') \in \mathcal{R}$.

En conjonction à (t), la propriété (i) induit une propriété clé des relations linéaires : la capacité de raisonner *ceteris paribus*– c.-à-d. toutes choses étant égales par ailleurs. La préférence de $\mathbf{y}$ sur $\mathbf{x}$ dépend uniquement de l'acceptabilité du *compromis* $\mathbf{y} - \mathbf{x}$, peu importe qu'il modifie $\mathbf{x}$ ou un autre $\mathbf{x}' \in \widehat{\mathbb{X}^n}$ (tant que $\mathbf{y}' := \mathbf{x}' + (\mathbf{y} - \mathbf{x})$ appartient à $\widehat{\mathbb{X}^n}$).

Définition 7 (Équivalence *ceteris paribus*)**.** Deux paires comparatives $(\mathbf{x}, \mathbf{y})$ et $(\mathbf{x}', \mathbf{y}')$ sont équivalentes *ceteris paribus* quand les alternatives $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$, appartiennent à $\widehat{\mathbb{X}^n}$ et les vecteurs $\mathbf{x} - \mathbf{y}$ et $\mathbf{x}' - \mathbf{y}'$ sont égaux. Dans ce cas, $(\mathbf{x}, \mathbf{y})$ et $(\mathbf{x}', \mathbf{y}')$ représentent le même *compromis*.

Quand une relation de préférence $\mathcal{R}$ vérifie (t) et (i), elle est compatible à l'équivalence *ceteris paribus*, dans le sens où deux paires équivalentes sont soit toutes les deux dans $\mathcal{R}$, soit aucune. Ainsi, $\mathcal{R}$ peut être définie par l'ensemble des *compromis acceptables* $to(\mathcal{R}) := \{\widehat{\mathbf{x}} - \widehat{\mathbf{y}}, (\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) \in \mathcal{R}\}$.

Lemme 4 (Ben-Porath et Gilboa, 1995)**.** *Une relation binaire réflexive $\mathcal{R}$ satisfait (t), (s), (r), (m), (d), (c) et (i) ssi il existe un n-uplet $\boldsymbol{\omega}$ de réels non négatifs et décroissants tel que $\mathcal{R} = O_{\mathrm{WA}_{\{\boldsymbol{\omega}\}}}$.*

C'est un théorème de représentation très puissant, cependant ses propriétés sont contradictoire vis à vis de notre programme de représenter le raisonnement sous-jacent par des règles déductives.

4.3 Au-delà de la décisivité

La propriété (d) est peu satisfaisante pour deux raisons. D'un point de vue raisonnement, elle équivaut à introduire la règle du *tiers exclus* dans notre système formel, ce qui permet d'introduire les preuves par réfutation [13] qui sont un outil puissant de déduction. Cependant, cela ne s'aligne pas avec notre besoin d'explications intelligibles, qui semble mieux correspondre à la logique intuitionniste. Du point de vue de la représentation de préférences, être décisif entre en conflit avec le besoin de capturer une information nécessairement incomplète. Un bon indicateur de la fragilité des décisions prises sous cette condition est que cette propriété est la seule dans cet article à ne pas être stable par intersection. Afin de proposer graduellement une transition entre la dominance de Lorenz généralisée (obtenue quand $\Omega = \Omega^\emptyset$) et une relation décisive (obtenue quand Ω est un singleton), nous adoptons le paradigme de *l'apprentissage robuste des préférences* [21, 20]. Supposons que l'on ait accès à de *l'information préférentielle* $\mathcal{I}$ sous la forme d'une relation binaire sur des alternatives, c.-à-d. un ensemble de référence de paires comparatives qui doivent être vérifiées par la relation de préférences recherchée.

Propriété ($\mathbf{pi}^{\mathcal{I}}$). Avec $\mathcal{I}$ une relation binaire sur des alternatives, $\mathcal{R}$ est compatible avec *l'information préférentielle* $\mathcal{I}$ quand $\mathcal{R} \supseteq \mathcal{I}$.

Du point de vue de la déduction, la compatibilité avec l'information préférentielle est obtenue naturellement en considérant les paires de $\mathcal{I}$ comme évidentes. Du point de vue de la représentation numérique, nous devons considérer l'ensemble $\Omega^{\mathcal{I}}$ contenant exactement tous les paramètres tels que $O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$ est compatible avec $\mathcal{I}$.

Définition 8 (ensemble des paramètres compatibles)**.** Avec $\mathcal{I}$ une relation binaire sur des alternatives, soit $\Omega^{\mathcal{I}}$ l'ensemble de tous les n-uplets $\boldsymbol{\omega}$ non-négatifs, décroissants, non-nuls de la sphère unité de $\mathbb{R}^n$ t.q. $O_{\mathrm{WA}_{\{\boldsymbol{\omega}\}}} \supseteq \mathcal{I}$. Quand $\Omega^{\mathcal{I}} \neq \emptyset$, $\mathcal{I}$ est dite *consistante* (avec l'*OWA* équitable).

L'inférence sceptique est souvent confronté à des problèmes computationnels [18], mais dans notre cas elle reste polynomiale, vu que l'*OWA* est linéaire par rapport à $\boldsymbol{\omega}$.

Lemme 5 (inspiré par [21])**.** *Étant donné une relation binaire sur des alternatives $\mathcal{I}$, l'ensemble $\Omega^{\mathcal{I}}$ est le polytope de $\mathbb{R}^n$ défini par les contraintes linéaires sur la variable $\boldsymbol{\omega}$: $\boldsymbol{\omega}_i \geq 0$ pour tout $i \in [n]$; $\boldsymbol{\omega}_i - \boldsymbol{\omega}_{i+1} \geq 0$ pour tout $i \in [n-1]$; $\sum_{i\in[n]} \boldsymbol{\omega}_i = 1$ et $\sum_{i\in[n]} (\widehat{\mathbf{b}}_i - \widehat{\mathbf{a}}_i)\boldsymbol{\omega}_i \geq 0$ pour tout $(\mathbf{a}, \mathbf{b}) \in \mathcal{I}$. De plus, vérifier si $\mathcal{I}$ est consistante, ou si une paire comparative est dans $O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$ se réduit à un problème d'optimisation linéaire soluble en temps polynomial.*

Exemple 4. *Considérons l'information préférentielle $(\mathbf{b}, \mathbf{d})$ donnée par l'autorité centrale dans l'exemple 1. Le score d'OWA paramétré par $\boldsymbol{\omega} \in \Omega^{\mathcal{I}}$ donné à l'alternative $\mathbf{d}$ doit être plus quand que celui donné à $\mathbf{b}$. Ainsi $\widehat{\mathbf{d}} \cdot \boldsymbol{\omega} \geq \widehat{\mathbf{b}} \cdot \boldsymbol{\omega}$, ou de manière équivalente $(\widehat{\mathbf{d}} - \widehat{\mathbf{b}}) \cdot \boldsymbol{\omega} \geq 0$. Cette contrainte peut être interprétée comme le trade-off $(\widehat{\mathbf{d}} - \widehat{\mathbf{b}}) = (+4, -8, -10, +4, -2)$ étant désirable.*

4.4 Caractérisation des préférences d'un *OWA* robuste

Notre prochain objectif est de caractériser la structure de préférences induite par l'ensemble d'*OWA* à l'aide de propriétés actionables amenant à un moteur explicatif correct et complet. Nous aimerions garder (t), (s), (r), (m) et (i), mais retirer (d), (c). Nous avons déjà commenté (d), que nous souhaiterions remplacer par ($\mathrm{pi}^{\mathcal{I}}$), et bien que (c) soit un outil mathématique fantastique, elle est inutilisable du point de vue cognitif. Elle permet de prendre la limite

13. En effet, pour prouver que $\mathbf{y}$ est préférée à $\mathbf{x}$, il suffit de prouver que $\mathbf{x}$ ne peut pas être préféré à $\mathbf{y}$.

à gauche et à droite dans une séquence de paires comparatives mais comment définir de telles séquences, vérifier leur convergence et calculer leur limites. Il sera difficile pour un non-expert de comprendre et de commenter une telle propriété.

Cependant, comme nous le verrons plus tard, (t), (s), (r), (m), (c), (i) et (pi$^{\mathcal{I}}$) ne sont pas suffisantes pour caractériser les *OWA* équitables et robustes. C'est pourquoi nous introduisons une nouvelle notion, proche de celle de l'équivalence *ceteris paribus*, mais nettement plus puissante.

En partant de l'équivalence *ceteris paribus*, nous souhaitons incorporer la symétrie et relâcher l'égalité entre les compromis par l'existence d'un lien non négatif, nous amenant à la propriété plus forte suivante

Définition 9 (congruence entre paires comparatives)**.** Deux paires comparatives $(\mathbf{x}, \mathbf{y})$ et $(\mathbf{x}', \mathbf{y}')$ sont congruentes quand $\widehat{\mathbf{x}} - \widehat{\mathbf{y}}$ et $\widehat{\mathbf{x}}' - \widehat{\mathbf{y}}'$ sont non-négativement liées [14]. Dans ce cas, nous écrivons $(\mathbf{x}, \mathbf{y}) \equiv (\mathbf{x}', \mathbf{y}')$.

Propriété (cc). $\mathcal{R}$ est compatible avec la *congruence* quand, pour toutes alternatives $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$, si $(\mathbf{x}, \mathbf{y}) \equiv (\mathbf{x}', \mathbf{y}')$ alors $(\mathbf{x}, \mathbf{y}) \in \mathcal{R} \iff (\mathbf{x}', \mathbf{y}') \in \mathcal{R}$.

Il est à noter que la propriété (cc) implique (i) et (s). Cela joue un rôle important dans la caractérisation des *OWA* décisifs [15]. Nous pouvons maintenant nous demander si cette nouvelle structure des compromis acceptables peut être dérivée, ou est une conséquence logique, des propriétés (t), (s), (r), (m), (c), (i) et (pi$^{\mathcal{I}}$). La réponse à ces questions est non.

Théorème 4. *Quand $\Omega^{\mathcal{I}}$ n'est pas un singleton, la propriété (cc) ne peut être déduite de (t), (s), (r), (m), (c), (i) et (pi$^{\mathcal{I}}$).*

Contre-exemple. Soit $\Gamma := \{(z_1, z_2, z_3) \in \mathbb{R}^3$ vérifiant (1) $3z_1 + 2z_2 + z_3 \geq 3$ ou (2) $z_1 \geq 0$ et $z_1 + z_2 \geq 0$ et $z_1 + z_2 + z_3 \geq 0\}$, et $\mathcal{R}$ la relation binaire sur $\mathbb{R}^3$ définie par $(\mathbf{x}, \mathbf{y}) \in \mathcal{R}$ ssi $\widehat{\mathbf{y}} - \widehat{\mathbf{x}} \in \Gamma$. $\mathcal{R}$ satisfait (i) et (s) par construction. Elle est continue car l'ensemble Γ de compromis acceptables est fermé (comme l'union de l'intersection de préimages d'ensembles fermés par des fonctions continues). Elle est transitive car Γ est stable par rapport à l'addition (la somme de deux vecteurs vérifiant (1) ou (2) vérifient respectivement (1) or (2), et la somme d'un vérifiant (1) et de l'autre (2) vérifie (2)). (r) et (m) sont obtenues par la condition (2). Cependant, bien que le compromis $t_1 := (2, -4, 6)$ est acceptable (correspondant par exemple à la paire comparative $(0, 8, 10)$ contre $(2, 4, 16)$), le compromis $\frac{1}{2}t_1 = (1, -2, 3)$ ne l'est pas (alors qu'il correspond par exemple à la paire $(3, 8, 9)$ contre $(4, 6, 12)$). □

14. c.-à-.d au moins un des vecteurs est nul, ou il existe $\lambda > 0$ t.q. $(\widehat{\mathbf{x}} - \widehat{\mathbf{y}}) = \lambda(\widehat{\mathbf{x}}' - \widehat{\mathbf{y}}')$.

15. Une étape importante de la preuve établie l'équation quand le coefficient λ est l'inverse d'un entier positif, en raisonnant par transitivité et le tiers exclus.

Comme corollaire [16], les propriétés (t), (s), (r), (m), (c), (i) et (pi$^{\mathcal{I}}$) ne sont pas suffisantes pour caractériser les préférences d'*OWA* robuste $O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$. Nous introduisons une nouvelle règle d'inférence

$$\textbf{Règle CC :}\ \frac{(\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{d}) \equiv (\mathbf{a}, \mathbf{b})}{(\mathbf{c}, \mathbf{d})}\ \text{(compatibilité à la congruence)}$$

correspondant à la propriété (cc). Nous pouvons donc introduire notre résultat principal.

Théorème 5. *Pour toute information préférentielle $\mathcal{I}$ consistante avec l'OWA :*

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

Le théorème 5 caractérise sémantiquement et déductivement les préférences basées sur un *OWA* équitable et robuste : $O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$ est la seule relation binaire réflexive satisfaisant (t), (s), (r), (m), (cc) et (pi$^{\mathcal{I}}$); de plus $\mathbf{x}$ est moins préféré que $\mathbf{y}$ selon cette relation si, et seulement si, il existe une preuve établissant cette préférence en utilisant uniquement les règles déductives T et CC et les vérités de $\mathcal{T}$, $\mathcal{G}$ ou $\mathcal{I}$. Comme décrit en section 3 dans le cas de la dominance de Lorenz, la chaîne d'inclusions de la gauche vers la droite dénote la correction, c.-à-d. $cl_{T,CC}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I}) \subseteq \mathcal{P}_{(t,r,m,cc,pi^{\mathcal{I}})} \subseteq O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$ est structurellement valide car les propriétés mises en avant correspondent à celles de l'OWA robuste et des règles. Nous allons maintenant concevoir un moteur explicatif implémentant une restriction de $cl_{T,CC}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I})$ et *complet* par rapport à $O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$, ce qui cloturera la preuve du théorème 5.

4.5 Moteurs explicatifs

L'exemple 1 illustrait le cas d'une paire comparative expliquée par une ATX basée sur les vérités premières de $\mathcal{S}$, $\mathcal{G}, \mathcal{T}$ et de $cl_{CC}(\mathcal{I})$ –les compromis qui sont congruents à un de ceux présents dans l'information préférentielle– mais ce moteur explicatif n'est peut être pas complet par rapport à $O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$, et est clairement difficilement traitable computationnellement car la contrainte de rester dans $\widehat{\mathbb{X}^n}$ est difficile à satisfaire, surtout quand les deux alternatives comparées sont proches du bord.

Nous proposons donc de relâcher le besoin de chercher un chemin de $\widehat{\mathbf{x}}$ vers $\widehat{\mathbf{y}}$ en la recherche d'un chemin de $\mathbf{x}'$ vers $\mathbf{y}'$ avec $(\mathbf{x}', \mathbf{y}')$ congruent à $(\mathbf{x}, \mathbf{y})$.

Définition 10 (Moteur CTX)**.** Soit $\mathcal{R}$ une relation binaire sur $\widehat{\mathbb{X}^n}$. Une *explication congruente et transitive basée sur les vérités dans* $\mathcal{R}$ de longueur ℓ ($\mathcal{R}$-CTX$^{\ell}$) est une paire (s, c) où le *support* s est un ℓ-uplet de paires comparatives $s = ((\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^{\ell}, \mathbf{y}^{\ell})) \in cl_{CC}(\mathcal{R})^{\ell}$ et *l'affirmation* c est une paire comparative $c = (\mathbf{x}, \mathbf{y})$ vérifiant $(\mathbf{x}^1, \mathbf{y}^{\ell}) \equiv (\mathbf{x}, \mathbf{y})$ et, pour tout entier k entre 2 et ℓ, $\mathbf{x}^k = \mathbf{y}^{k-1}$.

16. Le contre-exemple se focalise sur la famille de compromis $\{\delta : \omega \cdot \delta \geq K\}$, avec $\omega = (3, 2, 1)$, mais peut être modifié pour incorporer toute information préférentielle consistante mais non décisive.

Comme elles n'utilisent que des règles déductives et des vérités premières, il est clair que les CTX sont correctes par rapport aux règles déductives T et CC. Nous établissons un autre résultat principal, la complétude du moteur à l'aide d'une preuve constructive qui peut être appliquée par un algorithme efficace. Il conclut la preuve du théorème 5 (l'inclusion dans l'autre sens est obtenue via la correction).

Théorème 6. $O_{WA_{\Omega^{\mathcal{I}}}} \subseteq \bigcup_{\ell=1}^{n+|\mathcal{I}|} \mathcal{E}(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I}\text{-}CTX^{\ell})$

Démonstration. Soit $(\mathbf{x}, \mathbf{y}) \in O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$. Par le lemme de Farkas appliqué au programme linéaire du Lemme 5, il existe des coefficients non-négatifs $\langle \lambda^*_{(\mathbf{a},\mathbf{b})} \rangle_{(\mathbf{a},\mathbf{b})\in\mathcal{I}}$, $\langle \mu^*_i \rangle_{i\in[n]}$ et $\langle \nu^*_{i,j} \rangle_{1\le i<j\le n}$ t.q.

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

Cette équation décompose additivement le compromis correspondant à la paire comparative en 3 termes, chacun étant la somme de compromis correspondant aux vérités premières de $cl_{CC}(\mathcal{I})$, $\mathcal{G}$ et $\mathcal{T}$. Pour chaque agent i, nous séparons cette équation en deux parties, une contenant la somme des valeurs positives Δ^+_i et l'autre des valeurs négatives Δ^-_i, t.q. $\widehat{\mathbf{y}}_i - \widehat{\mathbf{x}}_i = \Delta^+_i + \Delta^-_i$. Si nous définissons $\widehat{\mathbf{x}}'$ et $\widehat{\mathbf{y}}'$ t.q. $\forall i \in [n] : \widehat{\mathbf{x}}'_i = \widehat{\mathbf{x}}_i + \sum_{j=1}^{i} \Delta^+_{j-1} - \sum_{j=1}^{i} \Delta^-_j$ et $\widehat{\mathbf{y}}'_i = \widehat{\mathbf{y}}_i + \sum_{j=1}^{i} \Delta^+_{j-1} - \sum_{j=1}^{i} \Delta^-_j$, nous avons *(i)* $\widehat{\mathbf{y}}_i - \widehat{\mathbf{x}}_i = \widehat{\mathbf{y}}'_i - \widehat{\mathbf{x}}'_i$, donc $(\mathbf{x}, \mathbf{y})$ et $(\mathbf{x}', \mathbf{y}')$ sont congruents ; et *(ii)* $\forall i \in [n]\ \widehat{\mathbf{x}}'_i - \widehat{\mathbf{x}}'_{i-1} \ge \Delta^+_{i-1} - \Delta^-_i$, ainsi la séparation entre les critères permet $\widehat{\mathbf{x}}'_{i-1}$ d'être augmenté et $\widehat{\mathbf{x}}'_i$ d'être réduit par les compromis donnés par le certificat de Farkas tout en vérifiant $\widehat{\mathbf{x}}'_{i-1} \le \widehat{\mathbf{x}}'_i$. Ainsi il est possible d'appliquer dans n'importe quel ordre les compromis de $\mathcal{T}$, $\mathcal{G}$ et $cl_{CC}(\mathcal{I})$ décrits par $\langle \lambda^*_{(\mathbf{a},\mathbf{b})} \rangle_{(\mathbf{a},\mathbf{b})\in\mathcal{I}}$, $\langle \mu^*_i \rangle_{i\in[n]}$ et $\langle \nu^*_{i,j} \rangle_{1\le i<j\le n}$ pour construire la chaîne transitive entre $\mathbf{x}'$ et $\mathbf{y}'$. Il en résulte que l'on peut toujours trouver une $(\mathcal{T} \cup \mathcal{G} \cup \mathcal{I})$-CTX de taille au plus $|\mathcal{I}| + n$ en temps polynomial. □

L'utilisation d'un certificat d'infaisabilité pour obtenir une explication peut aussi être trouvé dans [23, 3]. Obtenir ce certificat avec les résultats de la dualité forte (ici, le lemme de Farkas) peut aussi être trouvé dans [25, 4, 5].

Exemple 5. *Nous proposons de calculer une CTX, tel qu'expliqué dans la preuve du Théorème 6 pour la paire comparative* $(\mathbf{a}, \mathbf{c})$*, avec* $\widehat{\mathbf{c}} = (22\ 23\ 34\ 76\ 82)$ *et* $\widehat{\mathbf{a}} = (16\ 31\ 51\ 70\ 83)$*. Le certificat de Farkas obtenu est :* $\lambda^*_{(\mathbf{b},\mathbf{d})} = 1.5$*,* $\mu^* = (0\ 0\ 2\ 0\ 2)$ *et un unique transfert redistributif* $\nu^*_{3,2} = 4$*. Nous pouvons nous assurer de sa validité en calculant* $\widehat{\mathbf{c}} - \widehat{\mathbf{a}} = (+6, -8, -17, +6, -1) = 1.5 * (+4, -8, -10, +4, -2) + (0, 0, +2, 0, +2) + (0, +4, -4, 0, 0)$*. Nous générons les alternatives* $\mathbf{x}'$ *et* $\mathbf{y}'$ *comme décrit dans la preuve. Pour l'agent 1, nous avons* $\Delta^+_1 = 6$ *et* $\Delta^-_1 = 0$*, ainsi* $\mathbf{y}'_1 = \widehat{\mathbf{c}}_1 - \Delta^-_1 = \widehat{\mathbf{c}}_1 = 22$ *et* $\mathbf{x}'_1 = \widehat{\mathbf{a}}_1 - \Delta^-_1 = \widehat{\mathbf{a}}_1 = 16$*. Pour l'agent 2, nous avons* $\Delta^+_2 = 4$ *(du transfert redistributif) et* $\Delta^-_2 = -12$ *(issu de la* $\mathcal{I}$*-congruence), ainsi* $\mathbf{y}'_2 = \widehat{\mathbf{c}}_2 + \Delta^+_1 - \Delta^-_2 - \Delta^-_1 = 23 + 6 - (-12) = 41$ *et* $\mathbf{x}'_2 = \widehat{\mathbf{a}}_2 + \Delta^+_1 - \Delta^-_2 - \Delta^-_1 = 31 + 6 - (-12) = 49$*. Pour l'agent agent 3, nous avons* $\Delta^+_3 = -15 - 4 = -19$ *(du transfert redistributif et de la* $\mathcal{I}$*-congruence) et* $\Delta^-_3 = 20$ *(du don), ainsi* $\mathbf{y}'_3 = \widehat{\mathbf{c}}_3 + \Delta^+_2 + \Delta^+_1 - \Delta^-_3 - \Delta^-_2 - \Delta^-_1 = 34 + 4 + 6 - (-19) - (-12) = 75$ *et* $\mathbf{x}'_3 = \widehat{\mathbf{a}}_3 + \Delta^+_2 + \Delta^+_1 - \Delta^-_3 - \Delta^-_2 - \Delta^-_1 = 51 + 4 + 6 - (-19) - (-12) = 92$*. Nous continuons pour les agents 4 et 5 et obtenons ainsi* $\mathbf{y}' = (22\ 41\ 75\ 119\ 134)$ *et* $\mathbf{x}' = (16\ 49\ 92\ 113\ 135)$*. Nous avons bien* $\mathbf{y}' - \mathbf{x}' = \widehat{\mathbf{c}} - \widehat{\mathbf{a}}$*, donc les deux paires sont congruentes, nous pouvons donc construire la CTX de longueur 3 à partir de la chaîne d'alternatives* $(\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4)$ *avec* $\mathbf{x}^1 := \mathbf{x}'$ *et* $\mathbf{x}^4 := \mathbf{y}'$*. Toutes les permutations du transfert redistributif, du don et de la PI-congruence sont possibles, si nous optons pour cet ordre-ci nous avons* $\mathbf{x}^2 := (16\ 53\ 88\ 113\ 135)$ *and* $\mathbf{x}^3 := (22\ 41\ 75\ 119\ 134)$.

Nous pouvons nous demander si une ATX, plutôt qu'une CTX, peut être trouvée pour expliquer une affirmation appartenant à $O_{\mathrm{WA}_{\Omega^{\mathcal{I}}}}$. C'est en effet possible sous certaines conditions. Soit $\mathbb{F} := \{\mathbf{z} \in \widehat{\mathbb{X}^n} : \exists i\ \mathbf{z}_i = 0 \text{ ou } \exists j \neq i\ \mathbf{z}_i = \mathbf{z}_j\}$. $\mathbb{F}$ est la frontière de $\widehat{\mathbb{X}^n}$.

Théorème 7. *Soit* $\mathcal{I}$ *une relation binaire sur des alternatives consistante avec l'OWA et* $(\mathbf{x}, \mathbf{y}) \in O_{WA_{\Omega^{\mathcal{I}}}}$*. Si :*

i. aucune alternative $\mathbf{x}$ *ou* $\mathbf{y}$ *n'appartient à* $\mathbb{F}$ *; ou*

ii. $\sup_{\omega\in\Omega^{\mathcal{I}}} \boldsymbol{\omega} \cdot (\widehat{\mathbf{y}} - \widehat{\mathbf{x}}) > 0$

alors il existe une $[\mathcal{T} \cup \mathcal{G} \cup cl_{CC}(\mathcal{I})]$*-ATX supportant la conclusion* $(\mathbf{x}, \mathbf{y})$.

Cependant, notre preuve réside sur la construction d'une ATX de taille non bornée. Nous supposons donc l'existence de paires comparatives qui ne sont pas explicables par une ATX.

5 Conclusion et perspectives

Nous avons définis un moteur explicatif correct et complet pour les *OWA* robustes en utilisant un modèle logique formel. Nous ne sommes pas les premiers à suivre une telle approche, analogue à celle initiée par [15] et étendues par [33, 10, 30, 11]. Nous nous assurons que nos explications sont narrativement et cognitivement acceptables en évitant l'utilisation de propriétés purement techniques (comme la continuité), sous la forme d'une explication étape-par-étape, enchaînant les arguments, de taille bornée [9].

Nous pensons que nos résultats comblent une importante lacune : les *OWA* sont souvent utilisées pour des problèmes d'équités en optimisation combinatoire [31, 26], choix social computationnel [2, 27], apprentissage de préférences ou par renforcement [14, 17], il est raisonnable de penser que, quand l'équité est un aspect important, nous souhaitons aussi pouvoir examiner les décisions obtenues et éviter autant que faire se peut les biais non contrôlés ou les instabilités dues au choix de paramètres spécifiques. En donnant

des explications prouvablement correctes et lisibles nous pallions ce besoin. Nos explications peuvent ainsi permettre l'évaluation de la régularité de la procédure et l'adéquation des décisions algorithmiques, ou même le recours [24, 16].

Du point de vue IA de confiance, la prochaine évolution serait de soumettre les explications à l'approbation des propriétés sous-jacentes, avec des *questions critiques* [37] comme "est-ce raisonnable d'être symétrique ? (c'est l'anniversaire de Charlie)" ou "est-on sûrs que les utilités s'expriment sur une échelle commune ?". Cela requerrait d'intégrer le moteur explicatif dans un agent dialectique capable de raisonnements non monotones. Du point de vue théorie de la décision, la prochaine étape serait de transférer notre approche à des modèles plus complexes comme l'intégrale de Choquet, requérant un gros travail axiomatique.

Références

[1] Silvia Angilella, Salvatore Greco, and Benedetto 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.

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

[3] Khaled Belahcène, Yann Chevaleyre, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Accountable approval sorting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 70–76. ijcai.org, 2018.

[4] Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82 :151–183, 2017. Number : 2 Publisher : Springer.

[5] Khaled Belahcène, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Comparing options with argument schemes powered by cancellation. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, pages 1537–1543, 2019.

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

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

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

[9] Ignace Bleukx, Jo Devriendt, Emilio Gamba, Bart Bogaerts, and Tias 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.

[10] Arthur Boixel, Ulle Endriss, and Ronald 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] Arthur Boixel, Ulle Endriss, and Oliviero Nardi. Displaying justifications for collective decisions. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 5892–5895, 2022.

[12] Sylvain Bouveret, Yann Chevaleyre, and Nicolas Maudet. Fair allocation of indivisible goods. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors, *Handbook of Computational Social Choice*, pages 284–310. Cambridge University Press, 2016.

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

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

[15] Olivier Cailloux and Ulle Endriss. Arguing about voting rules. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 287–295. ACM, 2016.

[16] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. Trustworthy AI. In Bertrand Braunschweig and Malik Ghallab, editors, *Reflections on Artificial Intelligence for Humanity*, volume 12600 of *Lecture Notes in Computer Science*, pages 13–39. Springer, 2021.

[17] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Two-sided fairness in rankings via

lorenz dominance. *Advances in Neural Information Processing Systems*, 34 :8596–8608, 2021.

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

[19] Michael R Garey and David S Johnson. Computers and intractability : a guide to the theory of np-hardness, 1979.

[20] Salvatore Greco, Benedetto Matarazzo, and Roman Slowinski. Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129(1) :1–47, 2001.

[21] Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Ordinal regression revisited : Multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191(2) :416–436, 2008.

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

[23] Ulrich Junker. QUICKXPLAIN : preferred explanations and relaxations for over-constrained problems. In Deborah L. McGuinness and George Ferguson, editors, *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence*, pages 167–172. AAAI Press / The MIT Press, 2004.

[24] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165(3) :633–705, 2017.

[25] Christophe Labreuche, Nicolas Maudet, and Wassila Ouerdane. Justifying dominating options when preferential information is incomplete. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 486–491, 2012.

[26] Julien Lesca, Michel Minoux, and Patrice Perny. The fair owa one-to-one assignment problem : Np-hardness and polynomial time special cases. *algorithmica*, 81 :98–123, 2019.

[27] Jing Wu Lian, Nicholas Mattei, Renee Noble, and Toby 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.

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

[29] Radu Marinescu, Debarun Bhattacharjya, Junkyu Lee, Fabio Cozman, and Alexander Gray. Credal marginal map. In *Annual Conference on Neural Information Processing Systems*, 2023.

[30] Oliviero Nardi, Arthur Boixel, and Ulle Endriss. A graph-based algorithm for the automated justification of collective decisions. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS*, pages 935–943. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022.

[31] Wlodzimierz 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.

[32] Patrice Perny, Olivier Spanjaard, and Louis-Xavier Storme. A decision-theoretic approach to robust optimization in multivalued graphs. *Ann. Oper. Res.*, 147(1) :317–341, 2006.

[33] Dominik Peters, Ariel D. Procaccia, Alexandros Psomas, and Zixin Zhou. Explainable voting. In *Advances in Neural Information Processing Systems NeurIPS*, 2020.

[34] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you ?" : Explaining the predictions of any classifier. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 97–101. The Association for Computational Linguistics.

[35] Teddy Seidenfeld, Mark J Schervish, and Joseph B Kadane. Coherent choice functions under uncertainty. *Synthese*, 172 :157–176, 2010.

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

[37] Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

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

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