European Journal of Operational Research 328 (2026) 232–245

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Decision Support

Designing and computing explanations for comparisons inferred from an additive value model

Manuel Amoussou [b], Khaled Belahcene [a], Nicolas Maudet [b], Vincent Mousseau [a], Wassila Ouerdane [a,*]

[a] Université Paris-Saclay, CentraleSupélec, MICS, Gif-Sur-Yvette, 91190, France
[b] Sorbonne Université, CNRS, LIP6, France

ARTICLE INFO

Keywords:
Multiple criteria analysis
Explaining recommendation
Additive preferences model
Mathematical programming

ABSTRACT

Many decision models are based on an additive representation of preferences. Recommendations obtained from such additive decision models are sometimes considered as self-evident. On the contrary, we claim that these recommendations deserve an explanation so as to be fully understood by the user/decision-maker and to foster her trust. We propose to explain a preference statement x preferred to y by decomposing this statement into simpler ones. Arguments in favor of x (Pros), and arguments in favor of y (Cons) are decomposed using a *covering scheme* in which each Con is covered by a Pro. We use a *decomposition language* in which elementary self-evident statements involve (*i*) one Pro against one Con, (*ii*) one pro against several Cons, or (*iii*) several Pros against one Con. We prove that computing such explanations is computationally difficult in case (*ii*) and (*iii*), and propose a mathematical programming formulation to solve it. Numerical experiments provide insights on the actual behavior of our algorithm. We also illustrate the usefulness of our approach in the context of multicriteria decision aid but also for machine learning approaches.

1. Introduction

We consider a situation in which objects evaluated on several dimensions are to be compared using an additive model. Given an additive model has been elicited, the resulting comparisons need to be explained to a user (the explainee) in an intuitive way (which should not require a technical view of the additive model). We start with a description of three different situations in which such a need for explanation arises.

Evaluating students: At the end of the academic year, the academic jury collects students' grades and wishes to rank students in terms of merit, to award prizes to the ten best students. Once the results are announced, the student ranked in the 11th position requests explanations why the student ranked 10 is considered better than her.

Selecting buses: A transportation company operates buses and wants to open a new premium line. It owns a fleet of buses and needs to select three of them for the new service. The buses are evaluated according to eight quantitative criteria that reflect their performance and technical parameters. Selecting the three adequate buses can therefore be modeled as a multi-criteria choice problem. An additive preference model has been elicited following a multicriteria decision aid process. The selection of the three best buses (a, b, and c) should be explained to the bus fleet manager. In particular, she would need to understand why selecting buses a, b, and c is better suited than other buses (therefore explaining several comparisons).

Cancer diagnosis: Radiological images are being considered to assess the plausibility of patients having a specific cancer. A dataset of such images has been annotated, and cancer predictor features/attributes have been evaluated by radiologists. From this information, a prediction of the probability of cancer for a given patient image was obtained using a supervised learning technique. A radiologist is using the system which predicts that patient *Alice* has a higher cancer probability than patient *Bob*. She wishes to understand why the system makes such a statement.

These three example illustrate a generic explanation problem in which an *explainee* (student ranked in the 11th position, the bus fleet manager, the radiologist) requires an explanation for the statement "$\mathbf{x}$ is better than $\mathbf{y}$". Hence, the *explanandum*, i.e, the phenomenon to be explained, is a comparative statement: the student ranked 10 is better than the one in the 11th position, *Alice* has a higher cancer probability than *Bob*. We will call the *explainer*, the (software) agent computing and providing the explanation, in our case, a recommender system.

Previous work, Belahcene et al. (2017), proposed to explain a statement $\mathbf{x} \succ \mathbf{y}$ by decomposing this statement in a sequence of

* Corresponding author.
E-mail address: wassila.ouerdane@centralesupelec.fr (W. Ouerdane).

https://doi.org/10.1016/j.ejor.2025.05.058
Received 22 September 2022; Accepted 30 May 2025
Available online 13 June 2025
0377-2217/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

preference-swaps, i.e, "simple" preference statements considered as self-evident. A preference swap states that criterion A (in favor of $\mathbf{x}$) weighs more than criterion B (in favor of $\mathbf{y}$). Hence, this work proposes to define an explanation as matching between the *pros* (set of criteria in favor of $\mathbf{x}$ in the comparison) and *cons* (set of criteria in favor of $\mathbf{y}$). Such an explanation can be seen as intuitive as it decomposes the statement $\mathbf{x} \succ \mathbf{y}$ into elementary preference statements involving only two criteria. However, it does not allow the explanation of all comparisons and is therefore incomplete.

Our aim in this paper is to follow the path proposed by Belahcene et al. (2017) in which an explanation will consist of a decomposition of a preference $\mathbf{x} \succ \mathbf{y}$ into simpler preference statements while being able to explain most (if not all) statements. To do so, we propose to enlarge the set of elementary preference statements. In addition to (1,1)-trade-offs (criterion A, in favor of $\mathbf{x}$, weighs more than criterion B in favor of $\mathbf{y}$), we will consider the two types of non-decomposable preference statements described below.

- $(1, m)$-trade-offs: criterion A, in favor of $\mathbf{x}$, weighs more than criteria $B_1, B_2, \ldots, B_m$ (in favor of $\mathbf{y}$) together,
- $(m,1)$-trade-offs: criteria $A_1, A_2, \ldots, A_m$, in favor of $\mathbf{x}$, weighs more together than criterion B in favor of $\mathbf{y}$.

$(1, m)$ and $(m,1)$-trade-offs are natural extensions of (1,1) preference statements in which criteria are added on one side of the balance only. We claim these statements are intuitive and have an intelligibility argument as they cannot be further decomposed into sub-statements. As a comparison, (2,2)-trade-offs could be further decomposed and can be viewed as more complex to assess, as it requires to weight comparatively two pairs of criteria/attributes.

Belahcene et al. (2017) computes (1,1) decomposition-based explanations solving a matching problem. However, when extending the available arguments to encompass either $(1, m)$ or $(m, 1)$-trade-offs, or both, finding explanations becomes a computationally difficult combinatorial problem, as we will prove in this paper. We, however, propose an efficient resolution procedure by reduction to integer linear programming, study its performance, and provide illustrations of its applicability in different contexts.

The paper is organized as follows. Section 2 introduces all notations and definitions required to adequately specify our problem: computing an explanation for a $x \succ y$ statement given an elicited additive model. Section 3 specifies the computation of explanations using a covering problem and studies its computational complexity. An effective resolution technique based on Integer Linear Optimization is proposed. We perform, in Section 4, an experimental analysis which provides insights on the behavior and performance of our algorithm. Sections 5 and 6 illustrate how our proposal can be used in two different settings: Multi-Criteria Decision Analysis and Machine Learning based regression, respectively. Section 7 extends our approach to more complex recommendations. Section 8 analyzes how our work relates to the literature, and the last section groups conclusions and further research directions.

2. Definitions

Preferences based on additive values. We assume we are given a number of alternatives described by a number of viewpoints. The set of alternatives is $\mathcal{X}$, the set of viewpoints is $\mathcal{N}$. The desirability of a given alternative is given by a value function $V : \mathcal{X} \to \mathbb{R}$, such that an alternative $\mathbf{x}$ is preferred to an alternative $\mathbf{y}$ when $V(\mathbf{x}) \geq V(\mathbf{y})$. Moreover, we assume this value function is additive with respect to the viewpoints.

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \mathbf{x} \succsim \mathbf{y} \iff \sum_{i \in \mathcal{N}} v_i(\mathbf{x}) \geq \sum_{i \in \mathcal{N}} v_i(\mathbf{y}) \tag{2.1}$$

Example 2.1. For postgraduate training, licensed doctors are matched to hospitals. This is a two-sided market, where doctors-in-training have preferences over hospitals, and hospitals rank-order applicants according to their grades obtained during their training at medical school. Xavier ($\mathbf{x}$) and Yvonne ($\mathbf{y}$) both apply to the same hospital, Oxbridge General Hospital (OGH). Their grades are given by Table 1. The attentive reader will observe that the inequality (2.2) below holds, i.e, $x \succsim y$ given the data provided in Table 1.

$$\underbrace{8 \times 74 + 7 \times 89 + 7 \times 74 + 6 \times 81 + 6 \times 68 + 5 \times 84 + 6 \times 79}_{x} \geq \underbrace{8 \times 74 + 7 \times 71 + 7 \times 84 + 6 \times 91 + 6 \times 77 + 5 \times 76 + 6 \times 73}_{y} \tag{2.2}$$

According to Eq. (2.2), value of alternative $\mathbf{x}$ is greater than the value of $\mathbf{y}$, so the OGH favors the candidacy $\mathbf{x}$ over $\mathbf{y}$. Following common sense, the additive value model in general, and in particular its simplest form of a weighted sum – here, the grades obtained by the applicants needs not be re-encoded through a value function – is considered *interpretable*. It means that, in order to support their choice, the OGH could reveal the general formula (2.1) governing the model and the precise values of the grades and weights populating it, i.e. Table 1, and either let the explainee perform the calculation or somehow guide them through Eq. (2.2).

A *preference statement* is simply an ordered pair of alternatives $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^2$. It is *valid* when $\mathbf{x} \succsim \mathbf{y}$. In the context of a preference statement, criteria can be partitioned between pros – those according to which $\mathbf{x}$ is more desirable than $\mathbf{y}$ – cons, and neutral.

Definition 2.1 (*Orientation of a Viewpoint in the Context of a Preference Statement*)**.** Given a comparative statement $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^2$, $\mathcal{N}$ is partitioned into three subsets:

- *Pro viewpoints*: $\texttt{pros}(\mathbf{x}, \mathbf{y}) := \{i \in \mathcal{N} : v_i(\mathbf{x}) > v_i(\mathbf{y})\}$;
- *Con viewpoints*: $\texttt{cons}(\mathbf{x}, \mathbf{y}) := \{i \in \mathcal{N} : v_i(\mathbf{x}) < v_i(\mathbf{y})\}$; and
- *Neutral viewpoints*: $\texttt{neutral}(\mathbf{x}, \mathbf{y}) := \{i \in \mathcal{N} : v_i(\mathbf{x}) = v_i(\mathbf{y})\}$.

Syntax of explanations. We are interested in modeling an explanation as a speech act combining several sub-arguments into a compelling demonstration about why $\mathbf{x}$ should be preferred to $\mathbf{y}$. We normatively restrict the sub-arguments employed to refer to viewpoints solely. The sub-arguments included in explanations will typically involve one or several pro viewpoints (contained in $\texttt{pros}(\mathbf{x}, \mathbf{y})$) and one or several Con viewpoints (contained in $\texttt{cons}(\mathbf{x}, \mathbf{y})$), and state that the pros outweigh the cons (see Example 2.2 for illustrations). Syntactically, these sub-arguments can be defined as follows.

Definition 2.2 (*Contextualized Trade-offs*)**.** In the *context* of a given preference statement $(x, y) \in \mathcal{X}^2$, given two integers n_p, n_c, the set of trade-offs n_p pros and n_c cons is $\mathcal{T}_{x,y}^{n_p,n_c} := \{(P, C) \in \texttt{pros}(x, y) \times \texttt{cons}(x, y) : |P| \leq n_p \text{ and } |C| \leq n_c\}$. We call such trade-offs (n_p, n_c)-trade-offs. Two trade-offs (P, C) and (P', C') are said to be *disjoint* when both $P \cap P' = \emptyset$ and $C \cap C' = \emptyset$. A trade-off (P, C) is said to be *aligned with preference* when $\sum_{i \in P \cup C} v_i(x_i) \geq \sum_{i \in P \cup C} v_i(y_i)$

Semantically, the contextualized trade-off (P, C) expresses that the subset P of reasons to favor $\mathbf{x}$ is stronger than the subset C of reasons to favor $\mathbf{y}$.

Example 2.2 (*Example 2.1 Continued*)**.** In order to assess the polarity of criteria and the validity of trade-offs in the context of the valid preference statement $(\mathbf{x}, \mathbf{y})$, we map each criterion $i \in \mathcal{N} := \{A, B, C, D, E, F, G\}$ to its *algebraic strength* $v_i(\mathbf{x}) - v_i(\mathbf{y})$ (see Table 2).

The orientation of criteria depends on the sign of their algebraic strength. Consequently, $\texttt{pros}(\mathbf{x}, \mathbf{y}) = \{B, F, G\}$, $\texttt{cons}(\mathbf{x}, \mathbf{y}) = \{C, D, E\}$ and $\texttt{neutral}(alt x, \mathbf{y}) = \{A\}$. The validity of a trade-off is given by the sign of the algebraic strength of its pros and cons. For instance:

Table 1
Grades obtained by medical students and weights assigned to the subjects.

Candidate	Anatomy	Biology	Chemistry	Diagnosis	Epidemiology	Forensic pathology	Genetics
xavier	74	89	74	81	68	84	79
yvonne	74	71	84	91	77	76	73
Weight	8	7	7	6	6	5	6

Table 2
Difference in value between alternatives **x** and **y** associated to viewpoints. Sign denotes polarity of the viewpoint (either pro or con **x**, magnitude denotes its strength.

Viewpoint	A	B	C	D	E	F	G
Algebraic strength	0	+126	−70	−60	−54	+40	+36

- $(\{F\},\{\})$ is a (1,0)-trade-off, valid in the context of $(\mathbf{x},\mathbf{y})$ because F is a pro criterion. It corresponds to a dominance statement, and can be interpreted in the light of the *linear* property of the preference structure: "everything else being equal, high values of F are preferred to low ones".
- $(\{B\},\{C\})$ is a (1,1)-trade-off, valid in the context of $(\mathbf{x},\mathbf{y})$ because $(+126)+(-70) > 0$. In Belahcene et al. (2017), such a trade-off is called a *preference swap*, with the following *ceteris paribus* interpretation: "everything else being equal, we are ready to accept to give away some value according to the viewpoint F (forensic pathology), from $\mathbf{y}_C = 85$ to $\mathbf{x}_C = 74$, so as to achieve a value of $\mathbf{x}_B = 89$ rather than $\mathbf{y}_B = 71$ according to the viewpoint B (biology)".
- $(\{B\},\{D,E\})$ is a (1,2)-trade-off. It is valid in the context of $(\mathbf{x},\mathbf{y})$ because $(+126)+(-60)+(-54) > 0$. It can be interpreted as "everything else being equal, reaching a higher grade of 89 instead of 71 in biology more than compensates getting lower grades in diagnosis (81 instead of 91) and epidemiology (68 instead of 77)".
- $(\{F,G\},\{C\})$ is a (2,1)-trade-off which is valid in the context of $(\mathbf{x},\mathbf{y})$ because $(+40)+(+36)+(-60) > 0$, and has the following interpretation "taken together, the increases of grades in forensic pathology (from 76 to 84) and genetics (from 73 to 79) offset the decrease in chemistry (from 84 to 74)".

Definition 2.3 (*Covering Explanation*)**.** Given a non-negative integer $\ell \in \mathbb{N}$, a preference statement $(x,y) \in \mathcal{X}^2$ and a nonempty set of types $\mathcal{L} = \{(n_p^1, n_c^1), \dots\} \subset \mathbb{N}^2$, a $\mathcal{L}$ - covering explanation E supporting (x,y) of length ℓ is a set $E = \{(P_1,C_1),\dots,(P_\ell,C_\ell)\}$ of pairwise disjoint trade-offs in $\bigcup_{(n_p,n_c)\in\mathcal{L}} \mathcal{T}_{x,y}^{n_p,n_c}$ of cardinality ℓ such that $\bigcup_{i=1}^{\ell} C_i = \texttt{cons}(x,y)$.

Example 2.3 (*Example* 2.2 *Continued*)**.** Let $\mathcal{L}$ be the set of types $\mathcal{L} := \{(1,2),(2,1)\}$. $\{(\{B\},\{D,E\}),(\{F,G\},\{C\})\}$ is a $\mathcal{L}$-covering explanation of length 2 supporting $(\mathbf{x},\mathbf{y})$. Another explanation of length 2 in the same language $\mathcal{L}$ is $\{(\{B\},\{C,E\}),(\{F,G\},\{D\})\}$. It is interesting to note that there does not exist any $\mathcal{L}$-covering explanation when $\mathcal{L} := \{(1,1)\}$.

Soundness. The syntactic requirements placed upon sub-arguments and their combination yield an explanation engine that is *sound* w.r.t. additive preference.

Proposition 1 (*Soundness*)**.** *Let $(u_i)_{i\in\mathcal{N}}$ marginal value function defining an additive value model, and $(\mathbf{x},\mathbf{y})$ a preference statement. If $(\mathbf{x},\mathbf{y})$ is supported by a covering explanation where every trade-off is aligned with preference, then it is valid.*

Indeed, while the explanandum $\sum_{i\in\mathcal{N}} v_i(\mathbf{x}) - \sum_{i\in\mathcal{N}} v_i(\mathbf{y}) \geq 0$ is expressed in terms of the *aggregate and compare* paradigm, the explanans $\sum_{k=1}^{\ell}\left(\sum_{i\in P_k\cup C_k}(v_i(\mathbf{x})-v_i(\mathbf{y}))\right) + \sum_{i\notin\bigcup_{k=1}^{\ell} P_i\cup C_i}(v_i(\mathbf{x})-v_i(\mathbf{y}))$ can be read in the *compare then aggregate* paradigm. *Compare:* each summand $\sum_{i\in P_k\cup C_k}(v_i(\mathbf{x}) - v_i(\mathbf{y}))$ is obviously positive as it represents a self-evident trade-off aligned with preference. The last summand $\sum_{i\notin\bigcup_{k=1}^{\ell} P_i\cup C_i}(v_i(\mathbf{x}) - v_i(\mathbf{y}))$ is non-negative because each viewpoint in the set $\mathcal{N}\setminus\bigcup_{k=1}^{\ell} P_i\cup C_i$ is either a pro or neutral. *Aggregate:* the outer sum is positive because the summands are unanimously positive.

Informally (see Fig. 1), a covering explanation amounts to rearrange the $2|\mathcal{N}|$ terms of the sum $\sum_{i\in\mathcal{N}} v_i(\mathbf{x}) - \sum_{i\in\mathcal{N}} v_i(\mathbf{y})$, first as the sum of the algebraic strength of the viewpoints $\sum_{i\in Features}(v_i(\mathbf{x}) - v_i(\mathbf{y}))$, then by grouping the viewpoints so that each sub-sum is obviously positive, because it corresponds to a valid trade-off.

Incompleteness. At this point, if we refer to Definition 2.2, *any* trade-off can be considered in a covering explanation. It seems relevant to restrict the trade-offs involved with respect to the number of Pros and the number of Cons. In particular, note that if no restriction is imposed to n_p and n_c, there exists a trivial covering explanation of length 1 for any statement $x \gtrsim y$, which amounts to stating that the set of Pros outweighs the set of Cons (which is more a tautology than an explanation).

Therefore, we will restrict the trade-offs that can intervene in explanations to the ones we consider as self-evident and intuitive for decision-makers. For example, Belahcene et al. (2017) limits explanations to trade-offs to those involving one Pro and one Con. However, such limitation in the language will induce incompleteness. For instance, it is obvious that there is no covering explanation based on (1,1)-trade-offs only for assertions $x \gtrsim y$ with more Cons than Pros. The choice of the language defining the trade-offs used in explanations is crucial: as limited as possible so as to guarantee intelligibility, but as wide as necessary to ensure expensiveness and thus keep incompleteness reasonably unusual.

In this paper we will restrict explanations to those involving $(1,m)$-trade-offs and $(m,1)$-trade-offs. Indeed, we believe that:

(i) trade-offs with exactly one pro, and any number of cons aptly captures the elementary notion of *strength* of a favorable argument overcoming a number of minor drawbacks;
(ii) trade-offs with exactly one con, and any number of pros aptly capture the elementary notion of *accrual* of favorable reasons overcoming a single drawback;
(iii) trade-offs with strictly more than one con and one pro are cognitively challenging because they express comparative statements about the *intensity of preference*.

It should be noted that (i) and (ii) are irreducible in the sense that they cannot be further split into simpler parts. In addition, we support the claim (iii) with a decision-theoretic argument. Consider the following archetypal situation, involving four viewpoints A, B, C and D, such that:

$$\texttt{pros}(\mathbf{x},\mathbf{y}) = \{A,D\};\ \texttt{cons}(\mathbf{x},\mathbf{y}) = \{B,C\} \tag{2.3}$$

$$v_A(\mathbf{x}) + v_B(\mathbf{x}) + v_C(\mathbf{x}) + v_D(\mathbf{x}) > v_A(\mathbf{y}) + v_B(\mathbf{y}) + v_C(\mathbf{y}) + v_D(\mathbf{y}) \tag{2.4}$$

$$v_A(\mathbf{x}) + v_B(\mathbf{x}) + v_C(\mathbf{x}) < v_A(\mathbf{y}) + v_B(\mathbf{y}) + v_C(\mathbf{y}) \tag{2.5}$$

$$v_A(\mathbf{x}) - v_A(\mathbf{y}) > v_B(\mathbf{x}) - v_B(\mathbf{y}) > v_C(\mathbf{x}) - v_C(\mathbf{y}) > v_D(\mathbf{x}) - v_D(\mathbf{y}) \tag{2.6}$$

These equations ensure that **x** is preferred to **y**, but the viewpoint A in favor of **x**, while stronger individually than the viewpoints B and C in favor of **y**, is not strong enough to overcome them simultaneously. Meanwhile, the viewpoint D is in favor of **x** but too weak to trade favorably against either B or C. Thus, the comparative statement $(\mathbf{x},\mathbf{y})$, while valid, is not supported by any $\mathcal{L}$-covering explanation when $\mathcal{L} =$

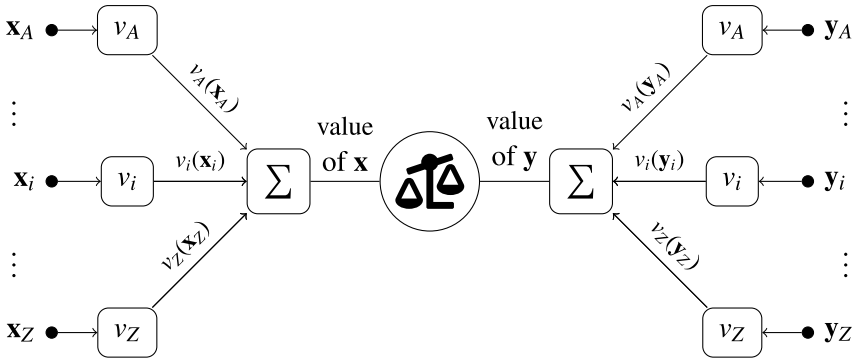

Fig. 1. The *explanandum*: $\mathbf{x}$ is preferred to $\mathbf{y}$, because its aggregated value is greater.

Table 3
Typology of comparative statements.

Denomination	# pros	# cons	Interpretation	Cognitive demand	Card.
$\mathcal{D}$	Any	None	Unanimous dominance statement	None	$2^m - 1$
$\mathcal{L}_0$	One	One	Preference swap	Light	$\Theta(m^2)$
$\mathcal{L}_1$	One	Any	Strong pro vs. a collection of minor cons	Moderate	$\Theta(m2^m)$
$\mathcal{L}_2$	Any	One	Accrual of pros overcoming a strong con	Moderate	$\Theta(m2^m)$
–	Two	Two	Cardinal statement wrt intensity of preference	High	$\Theta(m^4)$

$\{(1,2),(2,1)\}$. Thus, the explanation engine based on covering explanations is *not*, in general, *complete*—there are potentially valid statements that are left unexplained. Moreover, the trade-off $(\{A,D\},\{B,C\})$ can be understood as a statement concerning the intensity of preference: either as the preference for A over B being stronger than the preference for C over D, or the preference for A over C being stronger than the preference for B over D. This situation instantiates the quaternary relation "i is more intensely preferred to j than i' is preferred to j'", which is instrumental in the definition of *cardinal* values, while the types $(1,m)$ and $(m,1)$ remain on the arguably simpler side of *conjoint measurement* (von Winterfeldt & Edwards, 1986). The typology of trade-offs we consider self-evident or not is summarized in Table 3. These claims deserve to be further investigated from the perspective of cognitive sciences. In this paper, written from the viewpoint of computer science, we simply put them forward. In Section 3, we consider the computational problem of finding short $\mathcal{L}$-covering explanations for *any* set of self-evident trade-offs $\mathcal{L}$. In Sections 4–6, we investigate the expressive power of the covering explanation engine based on the sets of admissible trade-offs $\mathcal{L}_0 = \{(1,1)\}$, $\mathcal{L}_1 = \{(1,m)\}$, $\mathcal{L}_2 = \{(m,1)\}$, by either of those or by a mixture of those, i.e. $\mathcal{L}_3 := \{(1,m),(m,1)\}$.

3. Computing covering explanations

As stated in Definition 2.3, a covering explanation of length ℓ supporting a comparative statement $(\mathbf{x},\mathbf{y})$ is a partition of the viewpoints into ℓ subsets such as the preference for $\mathbf{x}$ over $\mathbf{y}$ when solely considering the viewpoints of each subset is self-evident. It can also be understood as a reordering of the $2|\mathcal{N}|$ terms appearing in the formula $\sum_{i\in\mathcal{N}} v_i(\mathbf{x}) - \sum_{i\in\mathcal{N}} v_i(\mathbf{x})$. Therefore, the problem of *computing* such an explanation can be seen as a search task over a combinatorial structure and is potentially arduous. In Section 3.1 we formalize and answer this question with a NP-completeness result as soon as subsets of viewpoints more expressive than mere *swaps* are allowed. In Section 3.2 we acknowledge this hardness result and propose an integer linear optimization (ILO) formulation allowing to use a generic solver to compute short explanations. In Appendix A, we provide an alternative, simpler formulation tailored to the case where $\{(1,1)\} \subsetneq \mathcal{L} \subseteq \{(1,p),(q,1)\}$ for some integers p,q.

3.1. Theoretical complexity

Problem statement. Let $\mathcal{L} := \{(p_1,q_1),\ldots\} \subset \mathbb{N}^2$ a finite, nonempty set of trade-offs. We consider the following function problem:

$\mathcal{L}$- Covering Explanation	
Input:	a finite, nonempty set $\mathcal{N}$ of viewpoints; two alternatives $\mathbf{x}$, $\mathbf{y}$ described by tuples $(v_i(\mathbf{x}))_{i\in\mathcal{N}}, (v_i(\mathbf{y}))_{i\in\mathcal{N}} \in \mathbb{N}^{\mathcal{N}}$; a positive integer L.
Output:	a $\mathcal{L}$ - covering explanation supporting $(\mathbf{x},\mathbf{y})$ of length $\ell \leq L$ where every trade-off is aligned with preference, if there is one.

We denote $\mathcal{L}$- Covering Explainability the corresponding decision problem, i.e. "is there a $\mathcal{L}$ - covering explanation supporting $(\mathbf{x},\mathbf{y})$ of length $\ell \leq L$ where every trade-off is aligned with preference?".

The tractable case $\mathcal{L} = \{(1,1)\}$. We observe that, in the case where trade-offs are restricted to the type $(1,1)$, i.e. *preference swaps*, the problem reduces to matching each con criterion with a pro argument of higher absolute value, as detailed in Belahcene et al. (2017). This matching problem can be solved in polynomial time with a dedicated algorithm. We provide here a novel operational characterization of explainability based on the notion of *one-dimensional walk*.

Proposition 2. *Given an instance* $\langle \mathcal{N}, (v_i(\mathbf{x}))_{i\in\mathcal{N}}, (v_i(\mathbf{y}))_{i\in\mathcal{N}}, L\rangle$ *of* $\{(1,1)\}$ −Covering Explanation, *define:*

- *the tuple* $\omega \in \mathbb{Z}^{\mathcal{N}}$ *such that* $\omega_i = v_i(\mathbf{x}) - v_i(\mathbf{y})$;
- *the permutation* $\pi : \mathbb{N} \to \mathcal{N}$ *such that* $|\omega_{\pi(1)}| \geq |\omega_{\pi(2)}| \geq \cdots \geq |\omega_{\pi(|\mathcal{N}|)}|$; *and*
- *the* $|\mathcal{N}|$*-tuple* ϵ *of elements in* $\{-1,+1\}$ *such that* $\epsilon_i = +1$ *if* $\omega_{\pi(i)} \geq 0$, -1 *else.*

The instance is positive if, and only if, $L \geq |\mathtt{cons}((\mathbf{x},\mathbf{y}))|$ *and* ϵ *is a walk on the line starting from the origin that remains non-negative, i.e. for all* $1 \leq k \leq |\mathcal{N}|$ $\sum_{i=1}^{k} \epsilon_i \geq 0$.

Proof. If the walk ϵ becomes negative at step k, then $\pi(k)$ is a con and $|\{i \in \mathtt{cons}((\mathbf{x},\mathbf{y})) : |\omega_i| \geq |\omega_k|\}| > |\{i \in \mathtt{pros}((\mathbf{x},\mathbf{y})) : |\omega_i| \geq |\omega_k|\}|$,

proving there is no matching of cons by stronger pros. Reciprocally, if there is a matching of cons by stronger pros, each component of ϵ equal to -1 is matched by a component of lower index equal to $+1$, yielding a non-negative walk. □

From the characterization provided by Proposition 2, we derive Algorithm 1, which consists in greedily associating the pro and con of highest magnitude while possible. The time complexity of Algorithm 1 is given by the sorting operation, hence $\Theta(|\mathcal{N}|\log|\mathcal{N}|)$.

Algorithm 1: Greedy algorithm for explaining with preference swaps

```
foreach i ∈ N do
    ω_i ← v_i(x) − v_i(y)
end foreach
π ← argsort(ω, key = |·|) ;          // π : ℕ → N such that |ω_π(1)| ≥ |ω_π(2)| ≥ ··· ≥ |ω_π(|N|)|
i ← 0 ; j ← 0 ; exp ← [] ;
while True do
    repeat    // finds next biggest con
        j ← j + 1 ;
        if j > |N| then return exp;    // no more con
    until ω_π(j) < 0 ;
    repeat    // finds next biggest pro
        i ← i + 1 ;
        if i > j then return False;    // pro cannot match con
    until ω_π(i) > 0 ;
    append ({i}, {j}) to exp ;
end while
```

Hardness of the general case. In the most general case, the syntactic constraints placed by the type are very loose – for instance, when the number of cons (resp. of pros) is left unbounded, an explanation is composed of arguments taken in a set with $\Theta(m2^m)$ elements – and we expect the problem of explainability to become difficult. Indeed, we prove the computational barrier comes even sooner than expected, as soon as pros or cons are allowed to combine.

Proposition 3. *When $\mathcal{L}$ contains an element (n_p, n_q) with $\max(n_p, n_q) > 1$, $\mathcal{L}$- Covering Explainability is NP-complete.*

Proof. Membership is straightforward, as a covering explanation is a partition of $\mathcal{N}$ and is thus a positive certificate of size $O(|\mathcal{N}|^2)$, while the input is in $\mathbb{N}^{2^{\mathcal{N}}}$.

Hardness results from reduction from Numerical 3D Matching (Garey & Johnson, 1979): given three multisets of integers A, B, C, each containing n elements, is there a subset $M \subset A \times B \times C$ such that every element in A, B and C occurs exactly once and each triplet $(a, b, c) \in M$ sums to $\sigma = \frac{1}{n}\sum_{x \in A \cup B \cup C} x$?

From an instance $A = \{a_1, \dots, a_n\}$, $B = \{b_1, \dots, b_n\}$, $C = \{c_1, \dots, c_n\}$ of Numerical 3D Matching, we build an instance of $\{(1,2)\}$ - Covering Explainability as follows. We assume w.l.o.g. $\sigma \geq 0$. Let $\mathcal{N} := [3n]$. Choose an integer m such that $b_j - m < 0$ for all $j \in [n]$. Choose an integer M large enough so that $a_i + m + M > 0$ for all $i \in [n]$, $c_k - M < 0$ for all $k \in [n]$, $a_i + b_j + b_k + M - m - \sigma > 0$ for all $i, j, k \in [n]$ and $a_i + c_j + c_k + m - 2M < 0$ for all $i, j, k \in [n]$. Define $\mathbf{x}$ such that $v_i(\mathbf{x}) = a_i + m + M$ if $1 \leq i \leq n$; $b_i - m - \frac{\sigma}{2}$, if $n+1 \leq i \leq 2n$; and $c_i - M - \frac{\sigma}{2}$ if $2n + 1 \leq i \leq 3n$. Let $\mathbf{y}$ such that $v_i(\mathbf{y}) = 0$ for all $i \in \mathcal{N}$. Observe $\sum_{i \in \mathcal{N}}(v_i(\mathbf{x}) - v_i(\mathbf{y})) = \sum_{i=1}^{n}(a_i + m + M + b_i - m - \frac{\sigma}{2} + c_i - M - \frac{\sigma}{2}) = 0$, $\texttt{pros}((\mathbf{x}, \mathbf{y})) = \{1, \dots, n\}$ and $\texttt{cons}((\mathbf{x}, \mathbf{y})) = \{n+1, \dots, 3n\}$. Suppose there is a numerical 3D matching M of A, B, C: each triplet a_i, b_j, c_k is associated to a subset of features $\{i, n+j, 2n+k\}$ containing one pro and two cons and such that $v_i(\mathbf{x}) + v_{n+j}(\mathbf{x}) + v_{2n+k}(\mathbf{x}) - (v_i(\mathbf{y}) + v_{n+j}(\mathbf{y}) + v_{2n+k}(\mathbf{y})) = a_i + b_j + c_k - \sigma = 0$ and are thus aligned with preference. Moreover these subsets of viewpoints are pairwise disjoint, thus constitute a valid $\{(1,2)\}$-covering explanation supporting $(\mathbf{x}, \mathbf{y})$.

Reciprocally, suppose there is a $\{(1,2)\}-$covering explanation supporting $(\mathbf{x}, \mathbf{y})$. The explanation is a null sum of non-negative terms, so each term is null. Moreover, each trade-off consists in one pro $1 \leq i \leq n$ and two cons $2n+1 \leq j < k \leq 3n$. Suppose $k \leq 2n$; the value associated with $\{i, j, k\}$ is $a_i + b_j + b_k + M - m - \sigma > 0$: a contradiction. Similarly, suppose $j \geq 2n+1$; the value associated with $\{i, j, k\}$ is $a_i + c_j + c_k + m - 2M - \sigma < 0$: a contradiction. Thus, $n+1 \leq j \leq 2n < 2n+1 \leq k \leq 3n$ and the value associated with $\{i, j, k\}$ is $a_i + b_j + c_k - \sigma$, thus $a_i + b_j + c_k = \sigma$. The con part $\{j, k\}$ of the trade-offs composing the explanation partition the cons, thus the explanation is of length n and the b_j anc c_k respectively partition B and C. Hence, there are n pros a_i that are pairwise disjoint, and thus cover A: the $\{a_i, b_j, c_k\}$ form a 3D numerical matching.

Similarly, it is possible to build a bijective function between positive instances of Numerical 3D Matching and positive instances of $\{(2,1)\}$ - Covering Explainability. □

3.2. Solving the general case with integer linear optimization

As the problem of finding short explanations is NP-complete, we propose to solve it using integer linear optimization (ILO), where integer variables are constrained by linear inequalities so as to minimize a linear cost function. Indeed, this type of formulation is popular in operations research and has received a lot of attention, which has fostered the development of several efficient solvers. More precisely, the formulation we give belongs to the *pseudo-boolean* fragment of ILO, where variables are restricted to $\{0, 1\}$.

The formulation is based on the explicit representation of each binary relations over viewpoints $\mathcal{R}_{(p,q)}$, where $(p, q) \in \mathcal{L}$ is an allowed type, such that $(i, j) \in \mathcal{R}_{(p,q)}$ iff the viewpoints i and j are associated in a trade-off of type (p, q) into a covering explanation. Each relation $\mathcal{R}_{(p,q)}$ is reflexive, transitive and symmetric by construction and thus is an equivalence relation. The trade-offs composing the covering explanation are thus the connected components (corresponding to equivalence classes, i.e. viewpoints appearing together) of these relations.

Pre-processing. Define $\mathcal{N}_* := \mathcal{N} \setminus \texttt{neutral}(\mathbf{x}, \mathbf{y})$. For all viewpoint $i \in \mathcal{N}_*$, define constants $\omega_i := v_i(\mathbf{x}) - v_i(\mathbf{y})$

Decision variables.

- a Boolean variable $a^{(\mathbf{x},\mathbf{y})}_{(p,q),i,j}$ for each type $(p, q) \in \mathcal{L}$, and pair of distinct viewpoints $i \neq j \in \mathcal{N}_*$, meaning viewpoints i and j belong together to a trade-off of type (p, q);
- a Boolean variable $b^{(\mathbf{x},\mathbf{y})}_{(p,q),k}$ for each type $(p, q) \in \mathcal{L}$ and viewpoint $k \in \texttt{pros}(\mathbf{x}, \mathbf{y}) \cup \texttt{cons}(\mathbf{x}, \mathbf{y})$, meaning the viewpoint k belongs to a trade-off of type (p, q);
- a Boolean variable $c^{(\mathbf{x},\mathbf{y})}_{(p,q),j}$ for each type $(p, q) \in \mathcal{L}$ and each con viewpoint $j \in \texttt{cons}(\mathbf{x}, \mathbf{y})$, meaning the viewpoint j belongs to some trade-off of type (p, q) and no con viewpoint of lesser index belongs to the same trade-off;
- a single Boolean variable $e^{(\mathbf{x},\mathbf{y})}$, meaning a covering explanation exists for the statement $(\mathbf{x}, \mathbf{y})$

Constraints.

- A viewpoint cannot appear in two distinct trade-offs.

$$b^{(\mathbf{x},\mathbf{y})}_{(p,q),i} \geq a^{(\mathbf{x},\mathbf{y})}_{(p,q),i,j} \text{ for all } i \neq j \in \mathcal{N}_*, \text{for all } (p, q) \in \mathcal{L} \tag{3.1a}$$

$$\sum_{(p,q) \in \mathcal{L}} b_{(p,q),i} \leq 1 \text{ for all } i \in \mathcal{N}_* \tag{3.1b}$$

- A trade-off of a given type (p,q) must refer to at most p pro viewpoints and q con viewpoints

$$\sum_{i\in \texttt{pros}(\mathbf{x},\mathbf{y})} a^{(\mathbf{x},\mathbf{y})}_{(p,q),i,j} \leq p \quad \text{for all} \quad j \in \texttt{cons}(\mathbf{x},\mathbf{y}) \quad \text{for all} \quad (p,q)\in\mathcal{L} \tag{3.2a}$$

$$\sum_{j\in \texttt{cons}(\mathbf{x},\mathbf{y})} a^{(\mathbf{x},\mathbf{y})}_{(p,q),i,j} \leq q \quad \text{for all} \quad i \in \texttt{pros}(\mathbf{x},\mathbf{y}) \quad \text{for all} \quad (p,q)\in\mathcal{L} \tag{3.2b}$$

- Reflexivity and transitivity of the relations $\mathcal{R}_{(p,q)}$

$$a^{(\mathbf{x},\mathbf{y})}_{(p,q),j,i} = a^{(\mathbf{x},\mathbf{y})}_{(p,q),i,j} \text{ for all } i\neq j\in\mathcal{N}_*, \text{for all } (p,q)\in\mathcal{L} \tag{3.3a}$$

$$1 + a^{(\mathbf{x},\mathbf{y})}_{(p,q),i,k} \geq a^{(\mathbf{x},\mathbf{y})}_{(p,q),i,j} + a^{(\mathbf{x},\mathbf{y})}_{(p,q),j,k} \text{ for all pairwise distinct } i,j,k\in\mathcal{N}_*, \text{for all } (p,q)\in\mathcal{L} \tag{3.3b}$$

- Trade-offs correspond to connected components and are indexed by their con viewpoint with the least index

$$\ell^{(\mathbf{x},\mathbf{y})}_{(p,q),j} + \sum_{j'\in\texttt{cons}(\mathbf{x},\mathbf{y}):\, j'<j} a^{(\mathbf{x},\mathbf{y})}_{(p,q),j,j'} \geq 1 \quad \text{for all} \quad j\in\texttt{cons}(\mathbf{x},\mathbf{y}) \quad \text{for all} \quad (p,q)\in\mathcal{L} \tag{3.4a}$$

- Each trade-off must be aligned with preference.

$$\omega_j + \sum_{i\neq j\in\mathcal{N}_*} \omega_i\, a^{(\mathbf{x},\mathbf{y})}_{i,j} \geq e^{(\mathbf{x},\mathbf{y})} - 1 \quad \text{for all} \quad j\in\texttt{cons}(\mathbf{x},\mathbf{y}) \tag{3.5}$$

Objective. The system of linear constraints above is always satisfiable. The comparative statement $(\mathbf{x},\mathbf{y})$ is supported by a covering explanation if, and only if

$$\max e^{(\mathbf{x},\mathbf{y})} = 1 \tag{3.6}$$

In such a case, an explanation can be constructed from the decision variables: each argument (P_j, C_j) corresponds to a connected component of the relation $\bigcup_{(p,q)\in\mathcal{L}} \mathcal{R}^{(\mathbf{x},\mathbf{y})}_{(p,q)}$; they can be reconstructed by Algorithm 2.

Algorithm 2: Retrieving a covering explanation from a solution of the ILO problem

```
assert e^(x,y) = 1 ;
I ← pros(x, y);
J ← cons(x, y);
exp ← empty list ;
while J ≠ ∅ do
    let j ∈ J;
    Ix ← ⋃_{(p,q)∈L} {i ∈ I : a^(x,y)_{(p,q)i,j} = 1} ;
    Jx ← ⋃_{(p,q)∈L} {j' ∈ J : a^(x,y)_{(p,q)j',j} = 1} ∪ {j} ;
    append (Ix, Jx) to exp;
    remove Ix from I;
    remove Jx from J;
end while
if I ≠ ∅ then
    append (I, ∅) to exp;
end if
return exp
```

Discriminating among explanations. As soon as the optimal value of the previous linear optimization problem is 1, any optimal solution yields an explanation. Yet, explanations are meant to serve as arguments in an interaction between a stakeholder of the decision and the decision maker. With this interaction between agents in mind, we might want to favor e.g. explanations that are *short* or employing sub-arguments that are *robust*. We formally define these requirements in terms of objectives of the linear optimization problem.

- Representing *length*: there is exactly one sub- argument of type (p,q) for each variable $l^{(\mathbf{x},\mathbf{y})}_{(p,q),j}$ equal to 1 in a given solution, thus the total number of sub-argument of type (p,q) is the sum over $j\in\texttt{cons}(\mathbf{x},\mathbf{y})$ of these variables, and the total number of sub-argument is obtained by summation over the types $(p,q)\in\mathcal{L}$. This total number (or a *weighted* combination of the number of sub-arguments of various types) can be added as a secondary objective, replacing Eq. (3.6) by the following

$$\max me^{(\mathbf{x},\mathbf{y})} + \sum_{(p,q)\in\mathcal{L}} \sum_{j\in\texttt{cons}(\mathbf{x},\mathbf{y})} \ell^{(\mathbf{x},\mathbf{y})}_j \tag{3.7}$$

- Representing *robustness*: the explainer might want to give precedence to sub-arguments (P,C) that correspond to comparative statements with a high intensity of preference, as measured by the value $\sum_{i\in P\cup C} w_i$. This value exactly corresponds to the positive slack of constraint (3.5). By representing this slack with a continuous variable $\sigma^{(\mathbf{x},\mathbf{y})}_j$, it is possible to maximize an aggregation (e.g. the sum or the minimum) of these variables.

As we advocate to restrain the set of acceptable arguments to the types one vs. any and any vs. one, i.e. consider $\mathcal{L}\subseteq\{(1,q),(p,1)\}$, we provide a simpler formulation for this specific situation in Appendix A.

4. Experimental assessment of the expressiveness of various argument sets

In this section, we empirically investigate, using synthetic data, the algorithm's behavior to compute explanations described in the previous section. In particular, we wish to assess the computing time required to compute an explanation for a pair $x \gtrsim y$, but also expressiveness of the explanation engines based on different sets of self-evident arguments: (1,1)-trade-offs, (1, m)-trade-offs, (m,1)-trade-offs, (1, m) and (m,1)-trade-offs.

4.1. Experimental protocol

The experimental design is the following: a data point is described by a tuple $(\omega_i)_{i\in\mathcal{N}}$, expressing the difference in marginal value between alternatives. Such tuples are sampled i.i.d. uniformly at random from $[-1,1]^{\mathcal{N}}$, with rejection when either (i) $\sum_{i\in\mathcal{N}}\omega_i < 0$ (i.e. the comparative statement is invalid); or (ii) $\forall i\in\mathcal{N}\ \omega_i \geq 0$ (i.e. the comparative statement is trivial, due to Pareto-dominance).

We consider a number of criteria/viewpoints varying between 3 and 40, so as to cover the usual range of MCDA applications (with typically less than 15 criteria), but also correspond to data-driven applications with a greater number of features and better observe trends. Once the number of criteria is fixed, we sample 10,000 datapoints, to ensure a 99% confidence interval demi-width of less than 0.01. Computations are made on a high-end 2023 laptop using Gurobi 12.0 to solve the integer linear optimization problems. We report the observed explainability rate in Table 4 and Fig. 2, the average length of the shortest explanation (when one exists) in Table 5 and Fig. 3, and computation time in Table 6 and Fig. 4.

Observe that the explainability rate for $\mathcal{L}_0$, i.e. the success of finding a $\{(1,1)\}$−covering explanation, can actually be computed by a closed formula.

Proposition 4. *The explainability rate for $\mathcal{L}_0$ when alternatives are sampled i.i.d. from a non-singular distribution with rejection when preference is invalid or trivial is given by $p(|\mathcal{N}|)$, with the function $p:\mathbb{N}\to[0,1]$ is:*

$$p(n) = \frac{\gamma(n+1)-2}{2^n-2} \text{ with } \gamma(2k) = \binom{2k}{k} \text{ and } \gamma(2k+1) = 2\binom{2k}{k}. \tag{4.1}$$

Table 4
Explainability according to the types of argument and the number of criteria.

nb. of criteria	3	4	5	6	7	8	9	10	11	12	13	15	20	25	30	40
$\mathcal{L}_0$	.67	.71	.60	.62	.54	.54	.49	.49	.45	.45	.42	.39	.36	.31	.30	.25
$\mathcal{L}_1$	.83	.77	.73	.69	.65	.62	.59	.57	.55	.53	.51	.48	.43	.38	.36	.31
$\mathcal{L}_2$	.83	.83	.80	.79	.77	.76	.75	.74	.73	.73	.71	.71	.69	.67	.67	.66
$\mathcal{L}_1$ or $\mathcal{L}_2$	1	.89	.92	.86	.87	.83	.84	.81	.82	.80	.80	.79	.75	.74	.73	.71
$\mathcal{L}_3$	1	.89	.92	.92	.91	.92	.92	.92	.93	.93	.93	.94	.95	.96	.97	.98

Table 5
Average explanation length according to the types of argument and the number of criteria.

nb. of viewpoints	3	4	5	6	7	8	9	10	11	12	13	15	20	25	30	40
$\mathcal{L}_0$	1.0	1.4	1.6	2.0	2.2	2.7	3.0	3.5	3.8	4.3	4.6	5.4	7.7	9.7	12	17
$\mathcal{L}_1$	1.0	1.1	1.3	1.5	1.7	1.9	2.1	2.4	2.6	2.9	3.1	3.6	4.9	6.3	7.6	11
$\mathcal{L}_2$	1.0	1.3	1.6	2.0	2.3	2.7	3.1	3.5	3.9	4.3	4.7	5.5	7.7	9.9	12	17
$\mathcal{L}_1$ or $\mathcal{L}_2$	1.0	1.1	1.4	1.5	1.9	2.1	2.4	2.7	3.1	3.3	3.7	4.4	6.1	8.1	9.9	14
$\mathcal{L}_3$	1.0	1.1	1.4	1.6	1.8	2.0	2.3	2.5	2.8	3.0	3.3	3.8	5.1	6.5	7.8	10

Proof. Let $|\mathcal{N}| = n$. Denote $\omega = (v_i(\mathbf{x} - v_i(\mathbf{y})))_{i\in\mathcal{N}}$ a random instance of $\{(1,1)\}$-Covering Explanation. Define the events E: ω is explainable; V: ω encodes a valid comparative statement $(\mathbf{x}, \mathbf{y})$; and T: ω encodes a trivial Pareto-dominance statement. Observe $E \subset V$ because explanations are sound, and obviously $T \subset E$. By definition, $p(n) = \mathbb{P}(E|V \text{ and not } T)$. By Bayes rule,

$$p(n) = \frac{\mathbb{P}(E \cap V \cap \bar{T})}{\mathbb{P}(V \cap \bar{T})} = \frac{\mathbb{P}(E) - \mathbb{P}(T)}{\mathbb{P}(V) - \mathbb{P}(T)}$$

Observe $\mathbb{P}(\mathbf{x} \succ \mathbf{y}) + \mathbb{P}(\mathbf{y} \succ \mathbf{x}) + \mathbb{P}(\mathbf{x} \sim \mathbf{y}) = 1$. The first two terms are equal to $\mathbb{P}(V)$ because $\mathbf{x}$ and $\mathbf{y}$ are i.i.d. The third term is null because the distribution is non-singular. Hence $\mathbb{P}(V) = \frac{1}{2}$. For the same reasons, each feature has a $1/2$ probability of being either a pro or a con. As they are sampled independently, $\mathbb{P}(T) = \frac{1}{2^n}$. Given $(|\omega_i|)_{i\in\mathcal{N}}$, Proposition 2 relates the occurrence of E to the configurations of polarities represented by a walk on a line starting from the origin and remaining on the non-negative side. Denote $\Delta(n)$ the set of such walks. It is related to $\Gamma(n)$, the set of n-step walks on a line starting from the origin but not returning to it: if $s \in \Delta(n)$ then appending a pro in first position yields an element of $\Gamma(n+1)$. Reciprocally, all the elements in $\Gamma(n+1)$ beginning with a pro remain on the positive side, thus removing this leading step yields an element of $\Delta(n)$. Thus $|\Delta(n)| = |\Gamma(n+1)|/2$. Thus $\mathbb{P}(E) = \frac{|\Gamma(n+1)|}{2^{n+1}}$ and

$$p(n) = \frac{\frac{|\Gamma(n+1)|}{2^{n+1}} - \frac{1}{2^n}}{\frac{1}{2} - \frac{1}{2^n}} = \frac{|\Gamma(n+1)| - 2}{2^n - 2}$$

To conclude, expressing the cardinality $\gamma(n)$ of the set $\Gamma(n)$ constitutes a classical problem (OEIS Foundation Inc., 2025, A063886). □

4.2. Results and discussion

The results of these experiments provide insights into the behavior of our algorithm on randomly generated data. Concerning computing time, although the problem tackled is computationally difficult (see Section 3), the actual time for computing explanations on datasets corresponding to real-world problem size is limited (see Table 6 and Fig. 4), and it seems compatible with its integration into an interactive process with a real user/decision-maker. The maximum observed execution time over all instances does not exceed one second (at most 0.6 s). This result is encouraging for the use of our algorithms on real-world instances. Indeed, these results provide an experimental guarantee that the explanation engine can be used in real-time with a decision-maker facing a real-world decision problem.

Fig. 2 and Table 4 depict for a varying number of criteria, and for the different languages, the proportion of preference pairs that can be explained in a given language. As expected, we observe the expressiveness of $(1,1)$-trade-offs – i.e. $\mathcal{L}_0$ – is limited and decreases fast, and that more expressive languages yield better explainability.

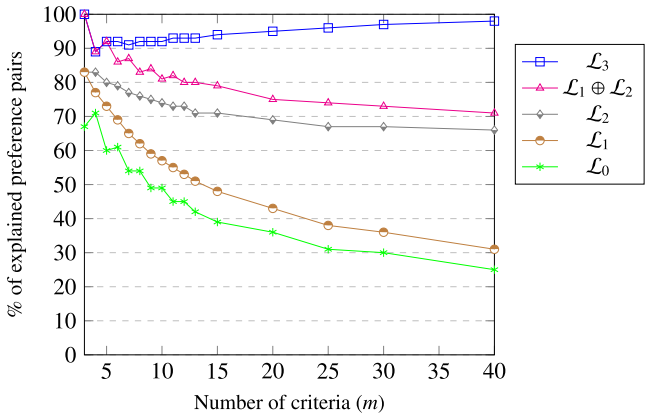

Fig. 2. Proportion of explainable pairs according to the set of types.

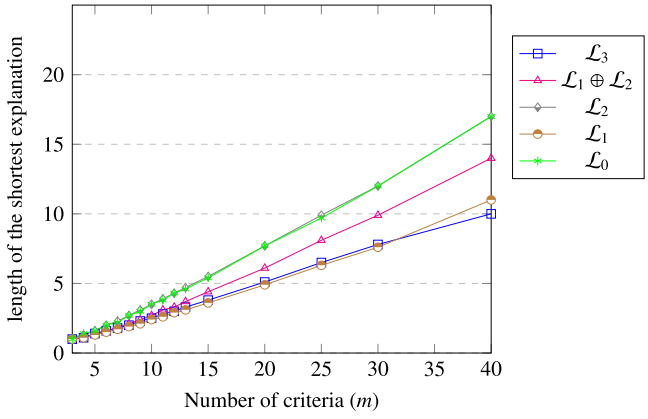

Fig. 3. Length of the shortest explanation according to the set of types.

Less expectedly, we observe that $\mathcal{L}_2$ allows to explain more preference pairs than $\mathcal{L}_1$, with a gap that increases with the number of criteria even though those languages obey symmetrical syntactic constraints. While allowing strong pro arguments to counterbalance an array of minor cons certainly helps at dealing with situations where cons are numerous but weak, allowing the strength of pros to accrue so as to overcome a single strong con unlocks many more situations. Moreover, we observe that explanations based on $\mathcal{L}_3$, where accrual and strength arguments can coexist, achieve much greater explainability than the others set of types and does not obey the same trend. Indeed, while every other explanation technique sees its efficiency decrease with the number of criteria, $\mathcal{L}_3$'s ability to explain expands, mildly but steadily, as the

Table 6
Average computation time (in milliseconds) according to the types of argument and the number of viewpoints.

nb. of viewpoints	3	4	5	6	7	8	9	10	11	12	13	15	20	25	30	40
$\mathcal{L}_1$	1.2	2.0	2.9	3.8	4.8	5.8	7.0	8.1	9.3	11	14	18	30	46	61	110
$\mathcal{L}_2$	1.1	1.5	2.0	2.8	3.7	4.7	5.7	6.9	7.8	8.9	11	14	22	33	51	120
$\mathcal{L}_3$	1.4	2.4	3.2	4.2	5.2	6.4	7.6	9.4	10.5	12	17	22	44	73	109	380

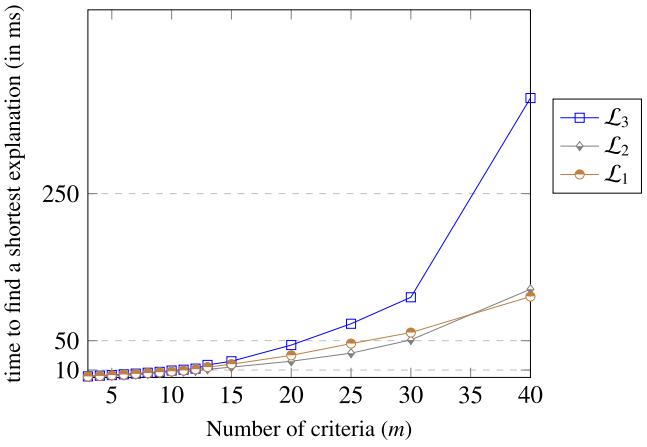

Fig. 4. Computation time according to the set of types.

number of criteria grows. It appears that $\mathcal{L}_3$, although still incomplete, is a language that makes it possible to explain "almost all" preference pairs.

Another important aspect of the experiment is the length of the computed explanation. Indeed, shorter explanations are simpler for the explainee (as there are less pieces of information to articulate) and, therefore, preferable. Table 5 and Fig. 3 depicts explanation length for each language, and for a varying number of criteria. With respect to explanation length, two languages outperform the others: $\mathcal{L}_1$, i.e. $(1, m)$-trade-offs, and $\mathcal{L}_3$, i.e. combining $(m,1)$-trade-offs and $(1, m)$. For these two languages, it should be emphasized that explanations remain relatively short: for a decision problem involving 10 criteria, explanations typically involve at most three arguments, which seems a "limited" cognitive burden for the explainee.

Overall, the results presented in this Section concerning the proportion of explained pair, explanation length, and computing time provide strong arguments in favor of the use of $\mathcal{L}_3$.

5. Application to a real-world choice problem

In this section, we show how our explanation engine can be tacked to a real world decision aiding context, in the case of a multiple criteria choice problem.

5.1. The context

A transportation company operates buses and wish to open a new premium line. It owns 76 buses and need to select three of them for the new service. The buses are evaluated according to eight quantitative criteria reflecting their performance and technical parameters. Selecting the adequate buses can therefore be modeled as a multicriteria choice problem.

The data. The criteria and performance table are taken from Żak and Stefanowski (1994) and has already served to illustrate the robust additive sorting model UTADIS^{GMS} in Greco et al. (2010). The criteria are presented in Table 7. The performance table for the 76 buses can be found in Table B.14 of Appendix B.

Solving the 3-best multicriteria choice problem is not the focus of this paper, and we use the well-known UTA method (Jacquet-Lagreze & Siskos, 1982) to learn piecewise linear additive function. We elected to use marginal value functions with 3 breakpoints. The specifics of the value function of the UTA model can be found in Table B.15 of Appendix B.

Table 7
Table of criteria.

Code	Name	Type
A	Maximum speed	Gain
B	Compression pressure	Gain
C	Blacking	Cost
D	Torque	Gain
E	Summer fuel consumption	Cost
F	Winter fuel consumption	Cost
G	Oil consumption	Cost
H	Horse power	Gain

Using the learned value function yields a score for each alternative (i.e. each bus) in $\mathcal{B} := \{\mathbf{b}_1, \ldots, \mathbf{b}_{76}\}$, and the three best are those of higher values: $\mathcal{B}^* = \{\mathbf{b}^*_{18}, \mathbf{b}^*_{29}, \mathbf{b}^*_{72}\}$.

5.2. Explaining the outcome

A first approach to explaining this outcome consists in trying to explain every statement in $\mathcal{B}_\star \times (\mathcal{B} \setminus \mathcal{B}_\star)$—there are $3 \times (76 - 3) = 219$ of them. Fortunately, most of these preference statements (209 out of 219) are dominance statements, where the first alternative is at least as good as the second one from all viewpoints, and strictly better from at least one viewpoint. Using the notation of Section 2, these statements are characterized by an empty `con` set and a nonempty `pro` set. They hold whatever the isotonic additive value function chosen, and do not require a sophisticated explanation engine—even though they fall under the umbrella of covering explanations of type (8,0). We now focus on the ten remaining statements:

$$\mathcal{E} = \{(\mathbf{b}^*_{29}, \mathbf{b}_{13}), (\mathbf{b}^*_{29}, \mathbf{b}_{61}), (\mathbf{b}^*_{29}, \mathbf{b}_9), (\mathbf{b}^*_{29}, \mathbf{b}_{49}), (\mathbf{b}^*_{29}, \mathbf{b}_1), (\mathbf{b}^*_{29}, \mathbf{b}_{55}), (\mathbf{b}^*_{72}, \mathbf{b}_{13}), (\mathbf{b}^*_{72}, \mathbf{b}_{61}), (\mathbf{b}^*_{72}, \mathbf{b}_9), (\mathbf{b}^*_{72}, \mathbf{b}_{49})\}$$

These statements are the edges of a graph whose vertices are $\mathcal{V} := \{\mathbf{b}^*_{29}, \mathbf{b}^*_{72}, \mathbf{b}_{13}, \mathbf{b}_{61}, \mathbf{b}_9, \mathbf{b}_{49}, \mathbf{b}_1, \mathbf{b}_{55}\}$. The attributes and corresponding value of the alternatives in $\mathcal{V}$ are presented in Table 8.

For each edge in $\mathcal{E}$, we determine the shortest $\mathcal{L}_3$-covering explanation, with $\mathcal{L}_3 = \{(1, 7), (7, 1)\}$ allowing to mix and match patterns of swaps, strong reason to accept and accrual of reasons to accept. These explanations are displayed in Table 9. Observe that the encoding of covering explanations ought to be understood *locally*, in the context of the explanandum. For instance, the explanans $(\mathtt{C}, \mathtt{DH})$ supporting the claim that $\mathbf{b}^*_{29}$ is preferred to $\mathbf{b}_{13}$ should be understood as:

"Everything else being equal, the quite lower blacking displayed by $\mathbf{b}^*_{29}$ *compared to* $\mathbf{b}_{13}$ *(18 vs. 26) is a strong reason to favor it, even though its torque and horse power are slightly lower (respectively 480 vs. 482 and 146 vs. 148)"*; whereas the explanans $(\mathtt{DF}, \mathtt{C}), (\mathtt{E}, \mathtt{GH})$ supporting the claim that $\mathbf{b}^*_{72}$ is preferred to $\mathbf{b}_{13}$ should be understood as:

"Everything else being equal,

- *together, the higher torque (484 vs. 482) and lower winter fuel consumption (23.6 vs. 24.5) displayed by* $\mathbf{b}^*_{72}$ *compared to* $\mathbf{b}_{13}$ *accrue to compensate a higher blacking (31 vs. 26); and*
- *the quite lower summer fuel consumption displayed by* $\mathbf{b}^*_{72}$ *compared to* $\mathbf{b}_{13}$ *(20.8 vs. 22.4) is a strong reason to favor it, even though its oil consumption is slightly higher (0.5 vs. 0.4) and its horse power is slightly lower (146 vs. 148)".*

Table 8
Performance and values of buses of interest, by decreasing order of preference. Chosen buses are starred.

Bus	A (v_A)	B (v_B)	C (v_C)	D (v_D)	E (v_E)	F (v_F)	G (v_G)	H (v_H)
$\mathbf{b}^*_{29}$	90 (.187)	2.58 (.095)	18 (.171)	480 (.178)	20.8 (.038)	23.4 (.121)	0.3 (.041)	146 (.118)
$\mathbf{b}^*_{72}$	90 (.187)	2.58 (.095)	31 (.136)	484 (.190)	20.9 (.038)	23.6 (.119)	0.5 (.040)	146 (.118)
$\mathbf{b}_{13}$	90 (.187)	2.58 (.095)	26 (.150)	482 (.184)	22.4 (.028)	24.5 (.106)	0.4 (.041)	148 (.121)
$\mathbf{b}_{61}$	90 (.187)	2.60 (.098)	34 (.128)	486 (.196)	21.1 (.036)	25.0 (.098)	0.6 (.039)	148 (.121)
$\mathbf{b}_{9}$	90 (.187)	2.56 (.091)	16 (.176)	486 (.196)	26.5 (.003)	27.3 (.065)	0.2 (.042)	150 (.124)
$\mathbf{b}_{49}$	90 (.187)	2.55 (.089)	38 (.118)	482 (.184)	20.8 (.038)	24.6 (.104)	0.7 (.039)	146 (.118)
$\mathbf{b}_{1}$	90 (.187)	2.52 (.084)	38 (.118)	481 (.181)	21.8 (.032)	26.4 (.078)	0.7 (.039)	145 (.116)
$\mathbf{b}_{55}$	90 (.187)	2.54 (.088)	47 (.094)	481 (.181)	22.0 (.030)	24.9 (.100)	1.0 (.037)	145 (.116)

Table 9
Shortest covering explanations for statements in $\mathcal{E}$.

vs	$\mathbf{b}_{13}$	$\mathbf{b}_{61}$	$\mathbf{b}_{9}$	$\mathbf{b}_{49}$	$\mathbf{b}_{1}$	$\mathbf{b}_{55}$
$\mathbf{b}^*_{29}$	(C,DH)	(C,BDH)	(F,CDGH)	(C,DG)	(C,D)	(C,D)
$\mathbf{b}^*_{72}$	(DF,C),(E,GH)	(F,BDH)	(F,C),(E,DGH)	(F,E)	No con	No con

Table 10
Predictors of breast cancer.

Code	Predictor	Weight
A	Clump thickness	.416
B	Uniformity Of cell size	.134
C	Uniformity Of cell shape	.247
D	Marginal adhesion	.235
E	Single epithelial cell size	.110
F	Bare nuclei	.352
G	Bland chromatin	.304
H	Normal nucleoli	.193
I	Mitoses	.226

Table 11
Values of predictors for two images.

Image	A	B	C	D	E	F	G	H	I
$\mathbf{x}_{522}$	10	4	3	2	3	10	5	3	2
$\mathbf{x}_{495}$	5	10	10	5	4	5	4	4	1

Explanations displayed in Table 9 answer queries concerning individual preference statements between a chosen alternative and a discarded one. Yet, if one were to explain as concisely as possible the entire sorting, i.e. why the starred alternatives are chosen and the unstarred ones are discarded, another approach is available, provided the explainer and explainee are ready to utter and hear, respectively, comparative statements between chosen alternatives. For instance, by explaining why $\mathbf{b}^*_{29}$ is preferred to $\mathbf{b}^*_{72}$ with a simple swap (C,D), the entire choice is supported with only seven explanantia (and transitive reasoning) instead of twelve. However, observe the full ranking $\mathbf{b}^*_{29} \succ \mathbf{b}^*_{72} \succ \mathbf{b}_{13} \succ \mathbf{b}_{61} \succ \mathbf{b}_{9} \succ \mathbf{b}_{49} \succ \mathbf{b}_{1} \succ \mathbf{b}_{55}$ cannot be explained with $\mathcal{L}_3$-covering explanations, because the comparative statements $(\mathbf{b}_9, \mathbf{b}_{49})$ and $(\mathbf{b}_1, \mathbf{b}_{55})$ are negative instances of $\mathcal{L}_3$-Covering Explainability.

6. Application to predictions made using logistic regression in the medical domain

In this section, we apply our explanation technique to a decision support system that, given information about cells located in the breast of female patients predicts the likelihood of the presence of a cancerous tumor. Images are described by a number of attributes detailed in Table 10 measured on ordinal scales ranging from 0 to 10.

Predictions are made using a model obtained using *logistic regression*, a simple yet widely-used supervised machine learning technique, on training data labeled by experts. Logistic regression is often used for binary classification tasks—predicting whether a data point belongs or not to a group of interest. However, this prediction actually relies on a prediction of the likelihood of membership, according to the following equation.

$$p = \frac{1}{1 + \exp(-\sum_{i \in \mathcal{N}} \omega_i \mathbf{x}_i)} \tag{6.1}$$

Moreover, approaches to learn a specific model from data have been designed that lead to *calibrated* values of the parameters, such that the probabilities obtained through Eq. (6.1) correspond to observations and predictions. Thus, likelihood of membership is monotonically increasing with the weighted sum of attributes, and we can use covering explanations to support comparative statements of the form "the sample collected from patient A is more likely to be benign than the one collected from patient B", that might in turn appear in deliberations among the medical staff on the way to prioritize further investigations or the necessity and urgency of potential interventions.

More specifically, we use a curated version of the publicly available dataset `breastcancer`,[1] where samples with missing data have been suppressed. We train a logistic regression model using the `scikit-learn` toolkit (Pedregosa et al., 2011) with cross-validation using default parameters, thus obtaining an 8-dimensional additive value model for p with 0.972 accuracy. The parameters of this model are reported in the last column of Table 10. We might have obtained a finer-grained model by using one-hot encoding on some or all features, so as to represent the fact that odd-ratio $\frac{p}{1-p}$ are not directly log-linear wrt the grades obtained on each scale.

Sample explanation. According to the model, the sample corresponding to image $\mathbf{x}_{522}$ is more likely to be benign than the one corresponding to image $\mathbf{x}_{495}$ (see Table 11). A $\mathcal{L}_1$-covering explanation of this fact is (F,BDE), (A,C), (G,H). We can produce a textual explanation leveraging the linearity of all marginal values, simply mentioning gains and losses without regard for the initial or final positions.

"Everything else being equal,

- *a gain of five units for bare nuclei more than offsets the losses of 6 points for uniformity of cell size, 3 for marginal adhesion and 1 for single epithelial cell size;*
- *a gain of 5 points for clump thickness is more important than a loss of 7 points for uniformity of cell shape; and*
- *a gain of one point for bland chromatin is more important than a loss of one point for normal nucleoli".*

[1] The original dataset is available on the UCI Machine Learning repository (Wolberg et al., 1993). https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original The curated dataset has already been used for benchmarking purposes in Ustun and Rudin (2016) and is available at https://github.com/ustunb/miplib2017-slim/tree/master/models/data.

Table 12
Measures of explainability.

Explanation device	#explainable	% explainable	Excluding dominance
Dominance	133,620	.578	0
$\mathcal{L}_0$-covering (swaps only)	207,250	.896	.753
$\mathcal{L}_1$-covering (strong pro only)	217,515	.940	.858
$\mathcal{L}_2$-covering (accrual of pros only)	218,520	.945	.869
either $\mathcal{L}_1$ or $\mathcal{L}_2$	228,200	.986	.968
$\mathcal{L}_3$-covering (mixture of strong pros and accrual)	229,360	.991	.980
Total number of explananda	231,356	1	

Table 13
Performance table.

Candidate	A	B	C	D	E	F	G	Score	Rank
x	74	89	74	81	68	84	79	3521	2
y	74	71	85	91	77	76	73	3510	3
z	75	89	74	81	68	84	79	3529	1
t	73	71	85	91	77	76	73	3502	4
Weight	8	7	7	6	6	5	6		

Measures of explainability. The dataset contains 683 data points. The learned linear regression model allows to totally pre-order them according to their predicted likelihood of being benign. Excluding ties, there are 231,356 strict comparative statements between any two alternatives. These explananda are fed to the $\mathcal{L}$-Covering Explainability device, for various sets $\mathcal{L}$ of self-evident arguments, similarly to the study made on synthetic data conducted in Section 4. Results are displayed in Table 12, with the first column detailing the explanation device, the second one giving the number of positive instances, the third one the fraction of positive instances over the total number of instances, and finally the fourth one reporting the fraction of non-trivial positive instances over the total number of non-trivial instances, excluding dominance statements. Thus, the proportions in the third column gives an idea of the actual proportion of real-world instances that shall be left unexplained, which is less than one percent when using the most powerful device. The proportions in the fourth column can be directly compared to the results obtained with synthetic data with the same number of feature/attribute in Section 4. We observe the explainability of real-world data is higher than the one estimated through synthetic data drawn uniformly and independently at random—certainly because the predictors are statistically correlated, leading to situations with a higher number of pros than with i.i.d. sampling of the attributes.

7. Beyond pairwise preference statements

In this section, we consider a setting where the decision in need of an explanation concerns not a single preference statement, but a set of those. More specifically, Section 7.1 details the case where we are given a *ranking* of alternatives, while Section 7.2 details the case of a *choice* task, where we need to justify why one or more alternatives were selected instead of others.

7.1. Explaining a complete ranking

Consider the following situation:

Example 7.1 (*Example* 2.1 *Continued*)**.** We are now given 4 candidates, detailed in Table 13.

Given the ranking $z \succ x \succ y \succ t$, we are actually given 6 statements: $z \succ x$, $z \succ y$, $z \succ t$, $x \succ y$, $x \succ t$, and $y \succ t$; for any of which we might be required to give an explanation. Nevertheless, a reasonable strategy can simply be to explain why an alternative is preferred to its successor, i.e., $z \succ x$, $x \succ y$, and $y \succ t$. Assuming preference is transitive (which is usually the case, and explicitly enforced in the additive value model belongs), this allows to deduce the missing statements.

From the above example, one can argue, based on transitivity, that explaining a ranking of n alternatives can be formulated as the explanation of $n-1$ comparisons alternative at rank i is preferred to alternative ranked $i+1$, $i = 1..n-1$). In other words, when explaining a ranking, leveraging the transitive property of preference reduces the number of preference statements requiring an explanation.

7.2. Explaining a choice

Combining explanations by covering and deduction with the transitive property can also be considered in the case of a choice problem

The 1-choice-explanation task. We need to explain why $x^\star$ is preferred to any other $y \neq x^\star$. A naive, immediate answer is to provide explanations supporting all the statements in $S := \{(\mathbf{x}^\star, \mathbf{y}) : \mathbf{y} \in \mathcal{X} \setminus \{\mathbf{x}^\star\}\}$. Unfortunately, this endeavor could fail for some preference statements. With this obstacle in mind, a reasonable approach consists, once again, in leveraging the transitive property. Instead of looking for a direct explanation of all statements $(\mathbf{x}^\star, \mathbf{y}), \mathbf{y} \neq \mathbf{x}^\star$, it would suffice to provide explanations of statements $\mathcal{E}$ forming a path from $\mathbf{x}^\star$ to each $\mathbf{y} \neq \mathbf{x}^\star$. Such arborescence rooted in $\mathbf{x}^\star$ is the equivalent in a directed graph of a spanning tree in an undirected graph. Given the full explainability graph $\mathcal{G} = (\mathcal{X}, E)$, where $(\mathbf{x}, \mathbf{y}) \in E$ iff this statement is $\mathcal{L}$- covering explainable, determining if there is such arborescence is easy by traversing the graph, and an efficient algorithm yielding a minimal one can be found in Gabow et al. (1986). Besides, observe that all minimal arborescence have the same size $|\mathcal{X}| - 1$: the transitive property helps finding explanations, but does not make the explanans shorter.

The k-best-explanation task. Suppose now the task we are given is to select a given number $k > 1$ alternatives among $\mathcal{X}$, then justify our choice. Formally, we now have a set $\mathcal{X}^\star$ of cardinality k, and we need to provide $\mathcal{L}$-covering explanations for all statements in a set $\mathcal{E}$ such that its transitive closure $cl(\mathcal{E})$ contains $\mathcal{X}^\star \times (\mathcal{X} \setminus \mathcal{X}^\star)$. A first approach consists in reducing this task to solving k 1-choice-explanation tasks and define $\mathcal{E} := \bigcup_{x^\star \in \mathcal{X}^\star} \mathcal{E}^{1-choice}_{x^\star, \mathcal{X} \setminus \mathcal{X}^\star}$. This approach yields an upper bound $|\mathcal{X}^\star|.|\mathcal{X} \setminus \mathcal{X}^\star|$ to the minimal size of $\mathcal{E}$. It might be far from optimal though (consider the favorable case where the Hasse diagram of the ranking of alternatives over $\mathcal{X} \setminus \mathcal{X}^\star$ is fully explainable). Computing the set $\mathcal{E}$ of explainable preference statements can be formulated using a mathematical programming approach, see Amoussou (2023).

Explaining the recommendations for the three above problem statements (Ranking, Choice, k-best) leads to defining broader problems. For each of these formulations, our proposal (explanation of a preference pair) can serve as a "building block" to compute relevant explanations for a more complex recommendation.

8. Related work

This contribution stands at the crossroads between three scientific fields: decision theory, knowledge representation and reasoning, and explainable artificial intelligence.

Decision theory is interested in providing normatively grounded models for describing agents' preferences and prescribing sound recommendations. In particular, the additive model has a long history.

Krantz et al. (1971) proposed a set of sufficient conditions allowing the representation of preferences with an additive function. Among these conditions, *first order cancellation* plays a key role in covering explanations because it warrants ceteris paribus reasoning. Keeney and Raiffa (1976) founded the multi-attribute value theory and contributed to the widespread use of this model to prescribe recommendations. Since then, the field of multicriteria decision aid has been further structured, with a wide array of methods and models allowing to address problems of varying nature (Bouyssou et al., 2006; Roy, 2013; Touskiàs, 2008). The pioneering work (Jacquet-Lagreze & Siskos, 1982) proposed an efficient way to learn an additive model from data. A thorough compilation of direct elicitation techniques can be found in von Winterfeldt and Edwards (1986), based on the notion of *primitive information* after which we model self-evident statements. Since then, many contributions in the field of multicriteria decision aid have proposed methodological papers where recommendation validation undergoes an elaborated process, including empirically grounded predictions and various assessments of their robustness (see e.g. Kadzinski et al., 2014; Kadzinski, Ghaderi et al., 2020). Several advances have been made to explain MCDA recommendation involving natural language processing (e.g. Papamichail & French, 2003; Wulf & Bertsch, 2017), argumentation (e.g. Labreuche et al., 2011), identifying preference premisses (e.g. Belanger & Martel, 2005; Kadziński, Badura et al., 2020), also in relation to Dominance-based Rough Set Approach (Greco et al., 2013).

Knowledge representation and reasoning (KRR) leverages explicit, declarative representation of knowledge. Of particular interest for the task of explanation are deductive, abductive and argumentative reasoning. Deduction is the process of applying given rules to given knowledge, and aptly describes the part of the explanans we propose where acknowledged arguments are aggregated together so as to match the explanandum. Abduction is the process of proposing pieces of information to obtain a given outcome by applying given rules, and aptly describes the fabrication of the arguments composing the explanans. However, the most popular abductive approach to explanation relies on the notion of *prime implicants*, minimally sufficient reasons to support a given conclusion (Darwiche & Hirth, 2023; Darwiche & Marquis, 2021; Ignatiev et al., 1998). In the context of preferences grounded on an additive model, an implicant is simply a coalition of viewpoints strong enough to overcome all other viewpoints even if they were maximally strong cons. This device offers valuable insight concerning the stability of the preference in various counterfactual scenarios. The explanans we put forward serves a different purpose: it guides the explainee so she can deduce the explanandum on her own, starting from prima facie information – here, comparative statements that are simultaneously aligned with the model and deemed self-evident–and following normatively sanctified steps – in the case of an additive preference model, by reasoning ceteris paribus. We also borrow some notions from argumentative reasoning, as the explanans we put forward is similar to an *argumentation scheme* (Walton, 1996) and the explanation process we envision is akin to a *dialogue game* (Black et al., 2021), see also (Carenini & Moore, 2006).

Some contributions are also situated at the crossroads of Decision theory and KRR. Tools have been proposed to support the outcome of a voting rule with a deductive proof in Boixel and Endriss (2020), Cailloux and Endriss (2016). Explaining devices have been proposed for *robust* decision models, that base their recommendation not on a single, precise value of a parameter, but on a set of those, implicitly defined by its compatibility with e.g. previous observations, whose volume reflects the imprecise knowledge about the preferences. Both (Belahcene et al., 2017, 2019) address the robust additive model. The former proposes to use an elicitation mechanism called *even-swaps* (Hammond et al., 1998) where viewpoints are considered in a pairwise fashion, trading one for the other, and serves as the foundation of the covering explanation. The latter is built on a slightly extended version of the normative property of *higher order cancellation* possessed by the additive model. Both use a given set of previous observations as the set of self-evident statements (restricted to swaps for Belahcene et al., 2019).

Explainable Artificial Intelligence (XAI) is concerned with providing either "interpretable" predictive models, or "interpretable" and faithful approximations of complex models. From this perspective, our journey begins where most end. Indeed, the additive value model is often considered to be *interpretable*, in the sense that its application is transparent and self evident. Indeed, popular feature attribution approaches in XAI, such as LIME (Ribeiro et al., 2016) or based on Shapley values (Chen et al., 2023) implicitly or explicitly use a multiattribute additive model as a surrogate. While we certainly agree that the additive model is one of the simplest multiattribute value model, this does not mean it is simple enough to be considered straightforward, transparent, or self-evident. On the contrary, its widespread use and preeminent role in XAI emphasize the need to honestly and carefully analyze the challenges in checking the general adequacy and procedural regularity of such a model.

We posit these challenges come from the *aggregate then compare* paradigm underlying the additive model prescribes to do precisely this: computing separately the value of each alternative, then comparing those values. In turn, this requires the disclosure of the marginal values associated to the attribute levels of each alternative. Taken in isolation, these numbers are meaningless, because values are measured on interval scales, on which the choice of origin and unit points are arbitrary. At this point, it seems reasonable to ground the explanation on the provision of the marginal value functions, i.e. the entire model. In XAI parlance, this amounts to a *global* explanation, allowing to understand the functional dependency between any input (here, preference queries) and the corresponding output of the decision support system. While global interpretability is desirable, it implies a degree of involvement and cognitive effort on behalf of the explainee, and a level of disclosure of the model on behalf of the explainer that seem unnecessary w.r.t. the task of explaining a single outcome.

9. Conclusion and perspectives

In this paper, we tackle the question of explaining a recommendation obtained from a given additive preference model. More precisely, we aim to compute an explanation supporting a preference statement $x \succsim y$ obtained with a given additive model. This explanation amounts at decomposing the statement $x \succsim y$ into simple ones. Are considered simple statements involving (i) one criterion in favor of x (one Pro) against one criterion in favor of y (one Con), (ii) one criterion in favor of x (1 Pro) against m criteria in favor of y (m Cons), or (iii) m criteria in favor of x (m Pros) against one criterion in favor of y (1 Con). We propose a covering-based explanation engine (in which each Cons is balanced by some pros), and a corresponding Integer Linear Programming resolution technique. Numerical experiments show that our approach is feasible with real-world problem size, and identify the additional descriptive ability of our approach with respect to previous work (Belahcene et al., 2017). Two illustrative examples (in Multi-Criteria Decision Analysis, and Machine Learning) Show that our proposal is relevant to provide explanations for a wide class of problems using additive preferences.

Our work opens avenues for further research. A first research question is related to our claim that (1,1) trade-offs, (1, m) trade-offs, and (m,1) trade-offs are simple ones, and simpler than (2,2) trade-offs. Although we provide strong formal arguments, it seems important to assess experimentally that it is the case from a behavioral point of view. Behavioral experiments involving users are needed in this perspective.

A second aspect to be explored relates to the structure of the explanation: it would be interesting to explore and formalize the contribution of transitivity, including to reveal intermediate levels on the attributes in the construction of explanations.

A third aspect to be further developed relates to the fact that the additive model, which is here considered as given, is often derived from

Table B.14
Complete performance table.

	A	B	C	D	E	F	G	H
$\mathbf{b}_1$	90	2.52	38	481	21.8	26.4	0.7	145
$\mathbf{b}_2$	76	2.11	70	420	22.0	25.5	2.7	110
$\mathbf{b}_3$	63	1.98	82	400	22.0	24.8	3.7	101
$\mathbf{b}_4$	90	2.48	49	477	21.9	25.1	1.0	138
$\mathbf{b}_5$	85	2.45	52	460	21.8	25.2	1.4	130
$\mathbf{b}_6$	72	2.2	73	425	23.1	27.4	2.8	112
$\mathbf{b}_7$	88	2.5	50	480	21.6	24.7	1.1	140
$\mathbf{b}_8$	87	2.48	56	465	22.8	27.6	1.4	135
$\mathbf{b}_9$	90	2.56	16	486	26.5	27.3	0.2	150
$\mathbf{b}_{10}$	60	1.95	95	400	23.3	24.8	4.4	96
$\mathbf{b}_{11}$	80	2.41	60	451	21.7	26.1	1.7	125
$\mathbf{b}_{12}$	78	2.4	63	448	21.8	26.0	1.9	120
$\mathbf{b}_{13}$	90	2.58	26	482	22.4	24.5	0.4	148
$\mathbf{b}_{14}$	62	1.98	93	400	22.0	28.4	3.9	100
$\mathbf{b}_{15}$	82	2.5	54	461	22.0	26.3	1.4	132
$\mathbf{b}_{16}$	65	2.22	67	402	22.0	23.9	2.3	103
$\mathbf{b}_{17}$	90	2.48	51	468	22.0	26.5	1.2	138
$\mathbf{b}^*_{18}$	90	2.6	15	488	20.0	23.2	0.1	150
$\mathbf{b}_{19}$	76	2.39	65	428	27.0	33.4	2.0	116
$\mathbf{b}_{20}$	85	2.42	50	454	21.5	26.3	1.3	129
$\mathbf{b}_{21}$	85	2.41	58	450	22.0	25.5	1.5	126
$\mathbf{b}_{22}$	88	2.47	48	458	22.4	25.1	1.1	130
$\mathbf{b}_{23}$	60	1.93	90	400	24.0	28.7	4.0	95
$\mathbf{b}_{24}$	64	2.2	71	420	23.1	25.2	2.6	105
$\mathbf{b}_{25}$	75	2.39	64	432	22.2	25.1	1.7	114
$\mathbf{b}_{26}$	74	2.36	64	420	21.9	25.4	1.9	110
$\mathbf{b}_{27}$	68	2.15	70	400	22.0	26.0	2.6	125
$\mathbf{b}_{28}$	70	2.2	65	412	22.8	25.3	2.1	102
$\mathbf{b}^*_{29}$	90	2.58	18	480	20.8	23.4	0.3	146
$\mathbf{b}_{30}$	83	2.39	64	445	22.1	26.0	2.0	120
$\mathbf{b}_{31}$	80	2.4	60	442	21.5	25.4	1.6	119
$\mathbf{b}_{32}$	88	2.49	44	478	21.8	25.2	0.9	138
$\mathbf{b}_{33}$	87	2.46	51	461	22.2	25.0	1.2	133
$\mathbf{b}_{34}$	85	2.4	55	445	23.0	26.2	1.5	120
$\mathbf{b}_{35}$	87	2.5	46	479	21.5	25.2	1.0	136
$\mathbf{b}_{36}$	85	2.41	65	450	22.0	26.1	1.9	125
$\mathbf{b}_{37}$	90	2.48	40	480	22.0	25.0	0.8	139
$\mathbf{b}_{38}$	72	2.36	64	428	21.9	25.4	2.0	111
$\mathbf{b}_{39}$	75	2.37	60	440	22.4	26.0	1.8	120
$\mathbf{b}_{40}$	75	2.36	68	441	23.1	26.0	2.1	113
$\mathbf{b}_{41}$	85	2.48	61	458	21.5	25.2	1.9	126
$\mathbf{b}_{42}$	86	2.48	52	462	22.1	24.8	1.5	129
$\mathbf{b}_{43}$	86	2.49	58	461	22.3	25.2	1.6	130
$\mathbf{b}_{44}$	88	2.5	48	475	22.0	24.5	1.1	140
$\mathbf{b}_{45}$	68	2.2	88	422	22.5	25.7	3.2	108
$\mathbf{b}_{46}$	72	2.36	78	424	22.4	25.5	2.8	112
$\mathbf{b}_{47}$	75	2.35	66	425	22.9	26.1	2.4	114
$\mathbf{b}_{48}$	82	2.38	65	430	23.0	25.8	2.3	115
$\mathbf{b}_{49}$	90	2.55	38	482	20.8	24.6	0.7	146
$\mathbf{b}_{50}$	84	2.36	64	438	23.3	26.2	2.1	119
$\mathbf{b}_{51}$	90	2.54	40	480	21.8	25.1	0.9	145
$\mathbf{b}_{52}$	89	2.51	46	470	22.0	25.5	1.1	141
$\mathbf{b}_{53}$	85	2.48	60	440	22.2	26.3	1.8	120
$\mathbf{b}_{54}$	90	2.52	45	479	21.6	25.1	1.0	145
$\mathbf{b}_{55}$	90	2.54	47	481	22.0	24.9	1.0	145
$\mathbf{b}_{56}$	85	2.5	60	446	21.9	24.6	1.9	123
$\mathbf{b}_{57}$	88	2.56	50	465	21.6	24.5	1.2	137
$\mathbf{b}_{58}$	84	2.32	65	428	22.0	25.1	2.2	116
$\mathbf{b}_{59}$	86	2.47	54	452	22.0	25.1	1.6	125
$\mathbf{b}_{60}$	66	2.29	86	402	23.3	26.6	3.2	105
$\mathbf{b}_{61}$	90	2.6	34	486	21.1	25.0	0.6	148
$\mathbf{b}_{62}$	67	2.24	86	406	23.5	27.0	3.0	104
$\mathbf{b}_{63}$	75	2.36	67	428	22.9	26.1	2.5	112
$\mathbf{b}_{64}$	86	2.42	60	444	22.0	25.2	1.8	122
$\mathbf{b}_{65}$	88	2.51	50	475	22.0	25.0	1.0	142
$\mathbf{b}_{66}$	85	2.38	63	440	21.8	26.0	2.1	120
$\mathbf{b}_{67}$	72	2.3	85	420	22.0	25.2	3.1	110
$\mathbf{b}_{68}$	75	2.4	69	428	20.9	25.6	2.4	115
$\mathbf{b}_{69}$	65	2.0	94	400	24.1	27.2	4.1	98
$\mathbf{b}_{70}$	87	2.4	60	460	22.0	25.6	1.7	131
$\mathbf{b}_{71}$	88	2.46	50	468	22.1	24.9	1.2	137
$\mathbf{b}^*_{72}$	90	2.58	31	484	20.9	23.6	0.5	146
$\mathbf{b}_{73}$	88	2.5	49	464	22.1	25.2	1.2	138
$\mathbf{b}_{74}$	87	2.48	52	465	21.9	24.6	1.4	135
$\mathbf{b}_{75}$	86	2.5	55	456	22.0	25.1	1.5	130
$\mathbf{b}_{76}$	88	2.52	46	472	21.8	23.8	1.1	141

Table B.15
Parameters of the UTA model.

A	B	C	D	E	F	G	H
30 ↦ 0	1.93 ↦ 0	15 ↦ .179	400 ↦ 0	20 ↦ .044	23.2 ↦ .124	0.1 ↦ .043	95 ↦ 0
60 ↦ .090	2.265 ↦ .040	55 ↦ .073	444 ↦ .072	23.5 ↦ .020	28.3 ↦ .051	2.25 ↦ .028	122 ↦ .081
90 ↦ .187	2.6 ↦ .098	95 ↦ 0	488 ↦ .202	27 ↦ 0	33.4 ↦ 0	4.4 ↦ 0	150 ↦ .124

a set of preference statements $\mathcal{PI}$. It would be interesting to consider a robust perspective in which any additive models compatible with $\mathcal{PI}$ could be used to derive explanations. Amoussou (2023) makes some proposals in this direction.

Lastly, our proposal can have an important impact in designing new interactive elicitation procedures. In such a procedure, the user, when presented with a provisional recommendation and its explanation, could reject preference information involved in the explanation during the dialogue. Additionally, one could consider learning an additive model under the constraint of explainability.

CRediT authorship contribution statement

Manuel Amoussou: Writing – review & editing, Methodology, Investigation, Conceptualization. **Khaled Belahcene:** Supervision. **Nicolas Maudet:** Supervision. **Vincent Mousseau:** Supervision. **Wassila Ouerdane:** Supervision, Funding acquisition.

Acknowledgments

This work has been supported by the project PSPC AIDA: 2019-PSPC-09 funded by BPI-France. Khaled Belahcène is supported by the chair "Explainable artificial intelligence for the future of Industry" funded by the ANR.

Appendix A. Computing $\mathcal{L}$-covering explanations when $\mathcal{L} \subseteq \{(1,q),(p,1)\}$

When all types $(p,q) \in \mathcal{L}$ satisfy $\min(p,q) = 1$, we propose a simpler and more efficient Integer Linear Optimization formulation for computing short $\mathcal{L}$-covering explanation, if one exists. This assumption corresponds precisely to the notion of self-evidence we put forward at the end of Section 2 summarized by Table 3. More specifically, we assume $\mathcal{L} = \{(p,1),(1,q)\}$. The case $\mathcal{L} = \{(p,1)\}$ is addressed by setting all variables u to zero. The case $\mathcal{L} = \{(1,q)\}$ is addressed by setting all variables v to zero.

Decision variables.

- a variable $u_{i,j}^{(\mathbf{x},\mathbf{y})}$ for each $i \in \texttt{pros}(\mathbf{x},\mathbf{y})$, $j \in \texttt{cons}(\mathbf{x},\mathbf{y})$ meaning viewpoints i and j are associated in a one vs. q (*strong pro*) argument;
- a variable $v_{i,j}^{(\mathbf{x},\mathbf{y})}$ for each $i \in \texttt{pros}(\mathbf{x},\mathbf{y})$, $j \in \texttt{cons}(\mathbf{x},\mathbf{y})$ meaning viewpoints i and j are associated in a p vs. one (*accrual of pros*) argument ;
- a variable $t_k^{(\mathbf{x},\mathbf{y})}$ for each $k \in \texttt{pros}(\mathbf{x},\mathbf{y}) \cup \texttt{cons}(\mathbf{x},\mathbf{y})$ meaning the viewpoint k is involved in one vs. q argument if it is a pro, or a p vs. one argument if it is a con;
- a variable $e^{(\mathbf{x},\mathbf{y})}$ meaning an explanation can be found for the comparative statement $(\mathbf{x},\mathbf{y})$.

Constraints.

- enforce semantics of t_i w.r.t. $u_{i,j}$ when i is a pro:

$$u_{i,j}^{(\mathbf{x},\mathbf{y})} \leq t_i^{(\mathbf{x},\mathbf{y})} \text{ for all } i \in \texttt{pros}(\mathbf{x},\mathbf{y}), j \in \texttt{cons}(\mathbf{x},\mathbf{y}) \tag{A.1}$$

- enforce semantics of t_j w.r.t. $v_{i,j}$ when j is a con:

$$v_{i,j}^{(\mathbf{x},\mathbf{y})} \leq w_j^{(\mathbf{x},\mathbf{y})} \text{ for all } i \in \texttt{pros}(\mathbf{x},\mathbf{y}), j \in \texttt{cons}(\mathbf{x},\mathbf{y}) \tag{A.2}$$

- a given pro viewpoint serves at most once:

$$\sum_{j \in \texttt{cons}(\mathbf{x},\mathbf{y})} v_{i,j}^{(\mathbf{x},\mathbf{y})} + t_i^{(\mathbf{x},\mathbf{y})} \leq 1 \text{ for all } i \in \texttt{pros}(\mathbf{x},\mathbf{y}) \tag{A.3}$$

- a given con viewpoint serves exactly once:

$$\sum_{i \in \texttt{pros}(\mathbf{x},\mathbf{y})} u_{i,j}^{(\mathbf{x},\mathbf{y})} + t_j^{(\mathbf{x},\mathbf{y})} = 1 \text{ for all } j \in \texttt{cons}(\mathbf{x},\mathbf{y}) \tag{A.4}$$

- at most q cons in each 1 vs. q argument:

$$\sum_{j \in \texttt{cons}(\mathbf{x},\mathbf{y})} u_{i,j}^{(\mathbf{x},\mathbf{y})} \leq q \text{ for all } i \in \texttt{pros}(\mathbf{x},\mathbf{y}) \tag{A.5}$$

- at most p pros in each p vs. 1 argument:

$$\sum_{j \in \texttt{cons}(\mathbf{x},\mathbf{y})} v_{i,j}^{(\mathbf{x},\mathbf{y})} \leq p \text{ for all } j \in \texttt{cons}(\mathbf{x},\mathbf{y}) \tag{A.6}$$

- alignment of each p vs. 1 arguments with preference:

$$1 - e^{(\mathbf{x},\mathbf{y})} + \omega_j + \sum_{i \in \texttt{pros}(\mathbf{x},\mathbf{y})} \omega_i v_{i,j}^{(\mathbf{x},\mathbf{y})} \geq t_j - 1 \text{ for all } j \in \texttt{cons}(\mathbf{x},\mathbf{y}) \tag{A.7}$$

- alignment of each 1 vs. q arguments with preference:

$$1 - e^{(\mathbf{x},\mathbf{y})} + \omega_i + \sum_{j \in \texttt{cons}(\mathbf{x},\mathbf{y})} \omega_j u_{i,j}^{(\mathbf{x},\mathbf{y})} \geq t_i - 1 \text{ for all } i \in \texttt{pros}(\mathbf{x},\mathbf{y}) \tag{A.8}$$

Objective. Maximize $me^{(\mathbf{x},\mathbf{y})} - \sum_{k \in \mathcal{N}_*} t_k$.

Appendix B. Details concerning real-world application of Section 5

The application concerns a fleet of buses. The data is taken from Żak and Stefanowski (1994). Each bus is described along eight viewpoints, detailed in Table 7. Table B.14 gives the performance of each bus. The alternatives playing a more detailed role in showcasing the explanation process are in gray. The UTA model is an additive value model $\mathcal{X} \to \mathbb{R}, \mathbf{x} \mapsto \sum_{i \in \mathcal{N}} v_i(\mathbf{x}_i)$ where each marginal function $\mathbf{x}_i \mapsto v_i(\mathbf{x}_i)$ is piecewise linear. We use 3 pieces evenly divided on the range of each attribute. These parameters are given by Table B.15.

References

Amoussou, M. (2023). *Explications en aide multicritère à la décision : schémas déductifs, algorithmes et expérimentations* (Ph.D. thesis), CentraleSupélec, Université Paris-Saclay.

Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., & Ouerdane, W. (2017). Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, *82*, 151–183.

Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., & Ouerdane, W. (2019). Comparing options with argument schemes powered by cancellation. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 1537–1543).

Belanger, M., & Martel, J. M. (2005). An automated explanation approach for a decision support system based on MCDA. In *Proceedings of the 2005 AAAI fall symposium*.

Black, E., Maudet, N., & Parsons, S. (2021). Argumentation-based dialogue. In Dov Gabbay, Massimiliano Giacomin, Guillermo R. Simari, & Matthias Thimm (Eds.), *Vol. 2, Handbook of formal argumentation*. College publication.

Boixel, A., & Endriss, U. (2020). Automated justification of collective decisions via constraint solving. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems* (pp. 168–176).

Bouyssou, D., Marchant, T., Pirlot, M., Tsoukias, A., & Vincke, P. (2006). *Vol. 86, Evaluation and decision models with multiple criteria: Stepping stones for the analyst*. Springer.

Cailloux, O., & Endriss, U. (2016). Arguing about voting rules. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems* (pp. 287–295).

Carenini, G., & Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence*, *170*(11), 925–952.

Chen, H., Covert, I. C., Lundberg, S. M., & Lee, S. (2023). Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, *5*, 590–601.

Darwiche, A., & Hirth, A. (2023). On the (complete) reasons behind decisions. *Journal of Logic, Language and Information*, *32*, 63–88.

Darwiche, A., & Marquis, P. (2021). On quantifying literals in Boolean logic and its applications to explainable AI. *Journal of Artificial Intelligence Research*, *72*, 285–328.

Gabow, H. N., Galil, Z., Spencer, T. H., & Tarjan, R. E. (1986). Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, *6*, 109–122.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability, a guide to the theory of NP-completeness*. Freeman.

Greco, S., Mousseau, V., & Słowiński, R. (2010). Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research*, *207*, 1455–1470.

Greco, S., Słowiński, R., & Zielniewicz, P. (2013). Putting dominance-based rough set approach and robust ordinal regression together. *Decision Support Systems*, *54*(2), 891–903.

Hammond, J., Keeney, R., & Raiffa, H. (1998). Even swaps: A rational method for making trade-offs. *Harvard Business Review*, *76*, 137–138, 143.

Ignatiev, A., Narodytska, N., & Marques-Silva, J. (1998). Abduction-based explanations for machine learning models. In *The thirty-third AAAI conference on artificial intelligence* (pp. 1511–1519). AAAI.

Jacquet-Lagreze, E., & Siskos, J. (1982). Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. *European Journal of Operational Research*, *10*, 151–164.

Kadziński, M., Badura, J., & Figueira, J. R. (2020). Using a segmenting description approach in multiple criteria decision aiding. *Expert Systems with Applications*, *147*.

Kadzinski, M., Corrente, S., Greco, S., & Słowiński, R. (2014). Preferential reducts and constructs in robust multiple criteria ranking and sorting. *Operations Research-Spektrum*, *36*, 1021–1053.

Kadzinski, M., Ghaderi, M., & Dabrowski, M. (2020). Contingent preference disaggregation model for multiple criteria sorting problem. *European Journal of Operational Research*, *281*, 369–387.

Keeney, R. L., & Raiffa, H. (1976). *DFecisions with multiple objectives: Preferences and value tradeoffs*. New York: J. Wiley.

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. i: Additive and polynomial representations*. New York: Academic Press.

Labreuche, C., Maudet, N., & Ouerdane, O. (2011). Minimal and complete explanations for critical multi-attribute decisions. In R. Brafman, F. Roberts, & A. Tsoukiàs (Eds.), *LNCS: vol. 6992, Algorithmic decision theory*. Springer.

OEIS Foundation Inc. (2025). The on-line encyclopedia of integer sequences. Published electronically at http://oeis.org.

Papamichail, K. N., & French, S. (2003). Explaining and justifying the advice of a decision support system: a natural language generation approach. *Expert Systems with Applications*, *24*(1), 35–48.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *TACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Roy, B. (2013). *Vol. 12, Multicriteria methodology for decision aiding*. Springer Science & Business Media.

Touskiàs, A. (2008). From decision theory to decision aiding methodology. *European Journal of Operational Research*, 138–161.

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*, 349–391.

Walton, D. (1996). Argumentation schemes for presumptive reasoning. *Cognitive Science*, *28*(5), 811–840.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge University Press.

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *Breast cancer wisconsin (diagnostic)*. UCI Machine Learning Repository.

Wulf, D., & Bertsch, V. (2017). A natural language generation approach to support understanding and traceability of multi-dimensional preferential sensitivity analysis in multi-criteria decision making. *Expert Systems with Applications*, *83*, 131–144.

Żak, J., & Stefanowski, J. (1994). Determining maintenance activities of motor vehicles using rough sets approach. In *Proc. of euromaintenance'94 conference*. Amsterdam.