Step-wise Explanations for the Additive Model

Manuel Amoussou, Khaled Belahcene, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane

Step-wise Explanations for the Additive Model

Manuel Amoussou
MICS, CentraleSupélec, Université Paris-Saclay
France
manuel.amoussou@centralesupelec.fr

Khaled Belahcène
Heudiasyc, Université de Technologie de Compiègne, CNRS
France
khaled.belahcene@hds.utc.fr

Nicolas Maudet
LIP6, Sorbonne Université, CNRS
France
nicolas.maudet@lip6.fr

Vincent Mousseau
MICS, CentraleSupélec, Université Paris-Saclay
France
vincent.mousseau@centralesupelec.fr

Wassila Ouerdane
MICS, CentraleSupélec, Université Paris-Saclay
France
wassila.ouerdane@centralesupelec.fr

ABSTRACT

We explore the problem of providing explanations for pairwise comparisons based on an underlying additive model. We follow a step-wise approach and provide explanations that take the form of a sequence of preference statements. Each statement should be as meaningful, relevant and cognitively simple as possible for the explanation to be accepted by an explainee. More specifically, we describe several schemes allowing to derive new knowledge, in the form of comparative statements, from previously accepted ones. These schemes exploit a number of well-understood properties of the additive model, and we ensure the correctness of the overall ex-planatory sequences. While these different schemes may correspond to alternative explanation strategies, we specifically advocate the use of the covering scheme because it meets some desirable properties for explanations. Imposing cognitively simple steps comes at the price of completeness. However, experimental results show that we are able to provide insightful explanations in many cases.

KEYWORDS

Additive Model, Explanation, Argument Schemes

1 INTRODUCTION

In this paper we address the problem of providing step-wise explanations for pairwise comparisons of alternatives based on well-established decision models. Alternatives are characterized by a number of criteria. The main idea is to break down the recommendation into simple statements presented to the explainee. The whole sequence of statements should formally support the recommendation. The explanations we aim for are thus *contrastive*, in the sense that the decision to be explained compares two alternatives, and *exact* (as opposed to *heuristic*) in the sense that we provide guarantees that the explanation produced is correct with respect to the underlying model. It is also common to distinguish between *local* explanations (when they focus on a specific recommendation) and *global* explanations (when they deal with the model in general): our approach is globally faithful to the model, and locally relevant to the pairwise comparison to be explained. Finally, even though this aspect is not detailed in this work, the perspective is to give the explainee the opportunity to accept or contradict these statements. To make things concrete we start with an illustrative example.

We consider seven abstract criteria ($\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g}$), each one described on bi-levels scales, which facilitate the symbolic representation of alternatives (*e.g* hotels). Each alternative can be represented as its evaluation vector ($s_1 = (✗, ✗, ✓, ✓, ✓, ✓, ✓)$) or more succinctly by the subset of criteria on which it is evaluated positively ($s_1 = \{\mathtt{cdefg}\}$). Moreover, for each criterion, the value symbolized by ✓ is more desirable than the value symbolized by ✗ (*e.g* breakfast included is better than not).

	a	b	c	d	e	f	g
s_1	✗	✗	✓	✓	✓	✓	✓
s_2	✓	✗	✗	✓	✗	✗	✗

The aggregation of criteria is done using an additive score function, assigning weights to the different criteria as follows:

$$\omega = \langle 128, 126, 77, 59, 52, 41, 37 \rangle$$

For example, the score of s_1 is thus equal to $w_{s_1} = 77 + 59 + 52 + 41 + 37 = 276$ while that of s_2 is: $w_{s_2} = 128 + 59 = 187$. It is also useful to encode the comparison of two alternatives as a vector $\{-1, 0, +1\}^n$ of arguments in favour (PRO) or against (CON) the option s_1, or neutral

(NEU). In our example, $\text{PRO}_{s_1} = \{\mathbf{c}, \mathbf{e}, \mathbf{f}, \mathbf{g}\}$, $\text{CON}_{s_1} = \{\mathbf{a}\}$, while $\text{NEU} = \{\mathbf{b}, \mathbf{d}\}$

Explanations can take many different forms. We list different possible explanations to the fact that s_1 is preferred to s_2:

(i) the first approach (*model disclosure*) could be to provide the full score calculation for both options, as illustrated above. But noticing that **d** is a neutral argument satisfied both by s_1 and s_2, we could omit it and simply provide the summation of PRO arguments vs. CON arguments.

(ii) the *counter-factual* approach seeks for minimal modification in the input that would change the outcome. For instance, we could state that, if s_2 had satisfied **b**, s_2 would instead have been recommended over s_1. Or (affecting the other alternative this time), if s_1 had not satisfied **cd**.

(iii) following a *prime implicant* approach, we could produce those arguments sufficient to explain the decision. In our case, two possible explanations could be given: (1) given that **bd** are neutral arguments, the PRO arguments **cef** are sufficient to overcome any set of CON arguments. In particular, this shows that the decision would remain the same even if **g** was a CON argument. And (2) given that **b** is a neutral argument, the PRO arguments **cefg** are sufficient to overcome any set of CON arguments. In particular, this shows that the decision would remain the same even if **d** was a CON argument.

(iv) following a *step-wise* approach, we could exhibit a collection of statements aiming at proving the decision. For instance, we could state that **cdefg** is preferred over **ac**, and that **ac** is preferred over **ad**, so that our conclusion should hold, following a transitive reasoning. Or, using a different logic, we could state that **cd** is preferred over **a**, while **efg** is preferred over **d**, which altogether justifies our decision.

This example allows us to illustrate some key principles of explanation (see *e.g.* [6, 17]) :

- *language intelligibility*—we want explanations to be conveyed in a language which is meaningful to the explainee. In our example, the weights used in calculations may not be easily interpreted by the explainee[1].
- *relevance*—we want explanations to focus on relevant information. In our example, as noticed, mentioning neutral arguments should be avoided if possible.
- *cognitive simplicity*—we want explanations to be "easy to process" by the explainee. This can be instantiated in different ways: prime implicant explanations are after (subset, or sometimes cardinality) minimal sufficient reasons, while step-wise explanations make use of intermediary comparisons involving a limited number of criteria.

Our ambition in this paper is to develop a principle-based and cognitively bounded model of step-wise explanations. As our example illustrates, there can be different 'logic' at play when combining statements. To account for that we describe a number of *schemes* for such explanations in the context of a comparison based on a weighted sum model (Section 3). By principle-based approach we mean that each scheme is attached to number of well-understood properties of the underlying decision model, that we make explicit and discuss in this paper. The resulting calculus is provably correct (Section 4). By cognitively-bounded we mean that our statements will be constrained so as to remain easy to grasp by the explainee. The resulting calculus is *not* complete, but we explore this issue in detail and provide several elements showing that our approach is satisfactory in terms of empirical completeness (Section 5).

[1]Another aspect, not investigated here, is that it may not be adequate to fully disclose the model for privacy or manipulation issues.

2 OUR MODEL

We consider a set of items $[m]$, and we abstractly refer to *states*, as subsets of items, i.e. elements of $2^{[m]}$. A *comparative statement* is a pair of states $(A, B) \in 2^{[m]} \times 2^{[m]}$, interpreted as a preference statement–'A is preferred to B'.

Schemes. Our aim is to provide a formal language and reasoning machinery allowing to support (explain) such comparative statements. We build on the notion of *argument scheme*, that is, an operator tying a sequence of statements, called the premise, satisfying some conditions, into another statement called the conclusion [20]. As we deal with preferences, argument schemes are ways of deriving new preferences from previously established ones. Noticeably, all our schemes operate on the same set of premises – finite sequences of comparative statements, represented as bracketed lists – and the same set of conclusions – comparative statements in $2^{[m]} \times 2^{[m]}$. We shall denote an arbitrary scheme s as:

$$[(A_1, B_1), \ldots, (A_k, B_k)] \xrightarrow{s} (A, B)$$

Correctness. The fact that our argument schemes allow us to only derive conclusions coherent with the preference relation is captured by the notion of correctness:

Definition 1. An argument scheme is *correct* w.r.t. a preference relation $\succsim$ if, when all premises belong to $\succsim$, then the conclusion also belong to $\succsim$.

At this stage we leave the preference relation unspecified, but in Section 3 we shall delve into this connection between the properties of preference relations and the schemes.

We further formalize the requirement of *relevance* (the absence of neural arguments) and *simplification* with respect to the relative difficulty of a statement.

Definition 2. A pair composed of a premise $[(A_1, B_1), \ldots, (A_k, B_k)]$ and a conclusion (A, B) is *independent of irrelevant alternative (III)* when $\left(\bigcup_{i=1}^{k} A_i \cup \bigcup_{i=1}^{k} B_i\right) \subseteq (A \cup B) \setminus (A \cap B)$.

Definition 3. A pair composed of a premise $[(A_1, B_1), \ldots (A_k, B_k)]$ and a conclusion (A, B) is *simplifying* when the premise is less difficult than the conclusion.

We believe this definition to be very general, as it captures one of the goals of explanation. To be actionable, though, it requires to specify the relative difficulty of a premise and a conclusion. We introduce a specific model allowing to derive the relative difficulty of statements, where this difficulty is purely syntactic and directly results from the number of items involved in the comparative statement.

Definition 4 (Difficulty of statements)**.** The *difficulty* of a comparative statement $(A, B) \in 2^{[m]} \times 2^{[m]}$ is the ordered pair of integers $(|A|, |B|)$. Consequently, we say that a comparative statement (A, B) is *less difficult than* another comparative statement (A', B')

when $|A| \leq |A'|$, $|B| \leq |B'|$ and at least one comparison is strict. A sequence of comparative statements $[(A_1, B_1), \ldots (A_k, B_k)]$ is *less difficult than* a comparative statement (A, B) when all comparative statements (A_i, B_i) are less difficult than (A, B). Finally, we define *difficulty classes* of comparative statements by putting upper bounds on the difficulty: for all integers p, q from 0 to m, let $\Delta(p, q) = \{(A, B) \in 2^{[m]} \times 2^{[m]} : |A| \leq p, |B| \leq q\}$.

We denote $\mathcal{A}$ the syntactically atomic elements, those that are considered self-evident and legit to be used as steps of an explanation for the considered explainee. We shall use difficulty classes $\Delta(p, q)$ to specify this set. In the context of explaining preferences between subset of desirable items, some values of the pair (p, q) are of specific interest: $\Delta(m, m)$ are unrestricted statements; comparative statements in $\Delta(m, 0)$ represent Pareto dominance statements; comparative statements in $\Delta(1, 1)$ can be interpreted as *swaps* [11], representing the exchange of one criterion against another; those in $\Delta(1, m)$ or in $\Delta(m, 1)$ represent a single item stronger or weaker than a subset of others, respectively considered as a pro or a con argument.

For instance, in the context of hotel comparisons, an argument in $\Delta(1, 1)$ could be *"we prefer to have free breakfast than free wifi access"*. An argument in $\Delta(1, 2)$ could be *"We prefer to have a swimming pool than free breakfast and wifi"*. To appreciate how difficult it can be to interpret higher order arguments, consider arguments in $\Delta(2, 2)$, for instance *"free breakfast and wifi access is preferable to having a swimming pool and being close to the city center"*. In Section 5 we shall investigate how restraining explanation to such simple statements affects the ability to produce explanations.

Explanation based on schemes. Self-evident atomic statements put a bound on the difficulty of each step of an explanation. As an explanation is a sequence of such statements, we also seek to produce correct explanations of minimal length:

Definition 5 (The explanation problem)**.** Given a comparative statement $(A, B) \in 2^{[m]} \times 2^{[m]}$, a preference relation $\succsim$, a set of statements $\mathcal{A}$ belonging to $\succsim$, a set of schemes $\mathcal{S}$, and a positive integer k: is there a positive integer $k' \leq k$, a list of length k' of statements $[(A_1, B_1), \ldots, (A_{k'}, B_{k'})]$ all belonging to $\mathcal{A}$, and a scheme $s \in \mathcal{S}$ such that $[(A_1, B_1), \ldots, (A_{k'}, B_{k'})] \xrightarrow{s} (A, B)$?

Note that this definition remains agnostic regarding the way the preference relation is represented in the input.

We shall now in the next section present the different argument schemes found in $\mathcal{S}$ and considered for reasoning with preferences.

3 SCHEMES FOR REASONING WITH PREFERENCES

This section is devoted to the construction of derivation rules adequate to reason about preferences. We formalize these rules as operators tying a list of premises to a conclusion, where premises and conclusions are comparative statements. We exploit key properties of the additive model–transitivity and cancellation–and formalize derivation rules taking advantage of each, from the ground up: the *transitive* and *ceteris paribus* schemes. We then introduce the *reduced transitive* scheme, that allows to directly derive any conclusion that can be proven using the two previous schemes. We conclude by introducing the *covering* scheme, which satisfies the independence of irrelevant items property (see Definition 2).

3.1 Properties of preference

We are interested in the preference relation $\succsim$ that might exist between states $A, B \in 2^{[m]}$, with $A \succsim B$ meaning that A is considered at least as good as B. We recall some useful features that preference relations may possess.

Definition 6 (Properties of preference)**.** Let $\succsim \subset 2^{[m]} \times 2^{[m]}$ a binary relation between states. We say:

- $\succsim$ is *transitive* when, for any states $A, B, C \in 2^{[m]}$, if $A \succsim B$ and $B \succsim C$ then $A \succsim C$;
- $\succsim$ satisfies (first order) *cancellation* if preference between states does not depend on common items, i.e. $\forall A, B \in 2^{[m]}\ A \succsim B \iff (A \setminus B) \succsim (B \setminus A)$;
- $\succsim$ is *additive* when there is a m-tuple of real numbers $\langle \omega_i \rangle_{i \in [m]} \in \mathbb{R}^{[m]}$ such that $A \succsim B \iff \sum_{i \in A} \omega_i \geq \sum_{i \in B} \omega_i$.
- $\succsim$ is an *additive linear order* when it is additive and there is no indifference, i.e. if $A \neq B$ then either $A \not\sim B$ or $B \not\sim A$ [9].

Obviously, an additive preference satisfies both the transitive and cancellation properties.

3.2 The transitive scheme

As we strive to explain recommendations deriving from a weighted sum model, we can mechanize the transitive and cancellation properties under the form of derivation rules. For instance, we define the *binary transitive scheme* $(2 - tr)$, allowing to chain preference statements:

$$[(A, B), (B, C)] \xrightarrow{2-tr} (A, C)$$

Following the approach we describe in Section 2, given an *explanandum* in the form of a preference statement (A, B) belonging to $\succsim$, an *explanans* is a proof consisting of recursive applications of a derivation rule – for instance, *2-tr* – allowing to derive the conclusion (A, B) from acceptable premises. Nevertheless, proofs are recursive objects that can be cumbersome to compute or present to an explainee, and we propose to alleviate this issue by introducing more powerful reasoning devices. Indeed, consider the case of purely transitive reasoning: chaining transitive lemmas amounts to consider chains of transitive premises. For instance, if we know that $A \succsim B$, $B \succsim C$, $C \succsim D$ and $D \succsim E$, we can infer that $A \succsim E$, which we denote $(A, B), (B, C), (C, D), (D, E) \vdash_{2-tr} (A, E)$. We believe this abundance of syntactic proofs not to be relevant to the question of computing explanations, and we therefore propose to consider the following *transitive scheme* (tr).

Definition 7 (transitive scheme (tr))**.** The premise $[(A_1, B_1), \ldots (A_k, B_k)]$ and conclusion (A, B) satisfy the *transitive scheme* when, for all $2 \leq j \leq k$, $A_j = B_{j-1}$, $A_1 = A$ and $B_k = B$.

Formally, $\vdash_{2-tr} = \xrightarrow{tr}$ – *what can be proven using the* $2-tr$ *scheme is exactly what can be derived in a single application of the* tr *scheme*. We have traded the recursive nature of the proof using a binary scheme for a *one-shot* derivation using a scheme operating on a list of premises of unbounded length.

EXAMPLE 1. *The premise* $[(\mathbf{acg}, \mathbf{bef}), (\mathbf{bef}, \mathbf{bfg})]$ *syntactically satisfies the transitive scheme for the conclusion* $(\mathbf{acg}, \mathbf{bfg})$*:*

$$[(\mathbf{acg}, \mathbf{bef}), (\mathbf{bef}, \mathbf{bfg})] \xrightarrow{tr} (\mathbf{acg}, \mathbf{bfg})$$

which can be expressed as: "as soon as $\mathbf{acg} \succ \mathbf{bef}$ *and* $\mathbf{bef} \succ \mathbf{bfg}$*,* $\mathbf{acg}$ *should be preferred to* $\mathbf{bfg}$*". Note however that the first comparative statement is complex as it involves six different criteria.*

Note that the comparative statements composing the premise of a transitive scheme are ordered, so the sequence of alternatives $A \equiv A_0 \succsim B_0 \equiv A_1 \succsim \cdots \succsim B_{k-1} \equiv A_k \succsim B_k \equiv B$ is non-increasing w.r.t. preference.

OBSERVATION 1. *If the preference* $\succsim$ *is transitive, then the transitive scheme is correct w.r.t.* $\succsim$

EXAMPLE 2. *In Example 1, the premise belongs to* $\succsim$*, as for* $(\mathbf{acg}, \mathbf{bef})$ *we have :* $242 = \omega_a + \omega_c + \omega_g > \omega_b + \omega_e + \omega_f = 219$*, and for* $(\mathbf{bef}, \mathbf{bfg})$*:* $219 = \omega_b + \omega_e + \omega_f > \omega_b + \omega_f + \omega_g = 204$*. However if we use the premise* $[(\mathbf{acg}, \mathbf{abc}), (\mathbf{abc}, \mathbf{bfg})]$*, this latter satisfies the transitive scheme but does not belong to* $\succsim$*, since for* $(\mathbf{acg}, \mathbf{abc})$*, we have:* $242 = \omega_a + \omega_c + \omega_g < \omega_a + \omega_b + \omega_c = 331$ *(see the score function* w *in Sect.2).*

3.3 The *ceteris paribus* scheme

The cancellation property allows to reason *ceteris paribus* – everything else being equal – independently from the context, and represents a great opportunity in terms of explanation, as preference can be deduced from comparative statements where common items are not mentioned. It motivates the following definition of the *ceteris paribus* argument scheme.

Definition 8 (ceteris paribus scheme (cp))**.** The premise $[(A_1, B_1), \ldots (A_k, B_k)]$ and conclusion (A, B) satisfy the *ceteris paribus scheme* when $k = 1$, $A_1 \setminus B_1 = A \setminus B$ and $B_1 \setminus A_1 = B \setminus A$. In this case, the comparative statements (A_1, B_1) and (A, B) are said to be *congruent*.

Obviously, congruence is an equivalence relation between comparative statements. In the congruence class of a given compartive statement $(A, B) \in 2^{[m]} \times 2^{[m]}$, the comparative statement $(A \setminus B, B \setminus A)$, where the states are pairwise disjoint and obtained from (A, B) by subtracting the common items $A \cap B$ respectively from A and B, is of special interest. When (A, B) and (A_1, B_1) are congruent, the preference of A over B translates to the preference of $A \setminus B$ over $B \setminus A$ by reasoning 'everything else' – in this case, items in $A \cap B$ – 'being equal' (ceteris paribus), and then to the preference of A_1 over B_1 by considering $A_1 \cap B_1$ irrelevant.

EXAMPLE 3. *When comparing* $\mathbf{acg}$ *to* $\mathbf{bfg}$*, it might be warranted to consider that, as they both achieve* $\mathbf{g}$*, this criterion can be omitted: 'the first option is better than the second one because, everything else being equal,* $\mathbf{ac}$ *is preferred to* $\mathbf{bf}$*'.*

Formally, we write $[(\mathbf{ac}, \mathbf{bf})] \xrightarrow{cp} (\mathbf{acg}, \mathbf{bfg})$

OBSERVATION 2. *If the preference* $\succsim$ *satisfies cancellation then the* ceteris paribus *scheme is correct w.r.t.* $\succsim$*.*

We shall often be presented with instances of the inverse problem: given an initial state, is there a final state such that the comparative statement from the initial to final state is congruent to a given comparative statement?

LEMMA 1 (FOURTH CONGRUENT PROBLEM). *Given a state* $A \in 2^{[m]}$ *and a comparative statement* $(A', B') \in 2^{[m]} \times 2^{[m]}$*, if* $A \supseteq (A' \setminus B')$ *and* $A \cap (B' \setminus A') = \emptyset$ *then there is exactly one state* $B \in 2^{[m]}$ *such that* (A, B) *and* (A', B') *are congruent, given by* $B = A \setminus (A' \setminus B') \cup (B' \setminus A')$*, else there is none.*

EXAMPLE 4. *Consider the comparative statement* $(\mathbf{ade}, \mathbf{bce})$*. The set of comparative statements congruent to it is:* $\{(\mathbf{ad}, \mathbf{bc}), (\mathbf{adf}, \mathbf{bcf}), (\mathbf{adg}, \mathbf{bcg}), (\mathbf{adef}, \mathbf{bcef}), (\mathbf{adeg}, \mathbf{bceg}), (\mathbf{adfg}, \mathbf{bcfg}), (\mathbf{adefg}, \mathbf{bcefg})\}$*. The initial states* $\{\mathbf{ad}, \mathbf{ade}, \mathbf{adf}, \mathbf{adg}, \mathbf{adef}, \mathbf{adeg}, \mathbf{adfg}, \mathbf{adefg}\}$ *are exactly the supersets of* $\mathbf{ad}$ *that do not contain* $\mathbf{bc}$*, and for each one of them, there is only one matching final state. This makes sense when the comparative statement* $(\mathbf{ade}, \mathbf{bce})$ *is understood as 'give* $\mathbf{ade}$*, take* $\mathbf{bce}$*', from which* $\mathbf{e}$ *can be omitted. Then,* $\mathbf{ad}$ *can only be effectively taken from a state already containing it, and* $\mathbf{bc}$ *can only be effectively added to a state not yet containing it.*

3.4 The reduced transitive scheme

When preference satisfies both the transitive and cancellation properties, it is correct to use both the tr and cp schemes to derive new comparative statements. Figure 1 illustrates such a proof.

$$\left.\begin{array}{l}(\mathbf{a}, \mathbf{b}) \xrightarrow{cp} (\mathbf{acg}, \mathbf{bcg}) \\ (\mathbf{c}, \mathbf{f}) \xrightarrow{cp} (\mathbf{bcg}, \mathbf{bfg})\end{array}\right\} \xrightarrow{tr} (\mathbf{acg}, \mathbf{bfg}) \xrightarrow{cp} (\mathbf{aceg}, \mathbf{befg})$$

Figure 1: A proof of $[(\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{f})] \vdash_{cp,tr} (\mathbf{aceg}, \mathbf{befg})$

For practical reasons, we want to streamline and mechanize this reasoning pattern, by formalizing a scheme–called *reduced transitive (rt)*–directly tying the premise $[(\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{f})]$ to the conclusion $(\mathbf{aceg}, \mathbf{befg})$ and leaving the intermediate steps unspecified[2].

Definition 9 (reduced transitive scheme (rt))**.** The premise $[(A_1, B_1), \ldots, (A_k, B_k)]$ and conclusion (A, B) satisfy the reduced transitive scheme when there exists $(A'_1, B'_1), \ldots, (A'_k, B'_k)$ and (A', B') such that:

$$\begin{cases} \forall i \in [k]\ [(A'_i, B'_i)] \xrightarrow{cp} (A_i, B_i); \\ [(A'_1, B'_1), \ldots, (A'_k, B'_k)] \xrightarrow{tr} (A', B'); \text{and} \\ [(A', B')] \xrightarrow{cp} (A, B). \end{cases}$$

EXAMPLE 5. *The proof depicted by Fig.1 can be read as follows: "*$\mathbf{acg}$ *should be preferred to* $\mathbf{bcg}$*, since everything else being equal,* $\mathbf{a}$ *is preferred to* $\mathbf{b}$*. Then* $\mathbf{bcg}$ *should be preferred to* $\mathbf{bfg}$*, since everything else being equal,* $\mathbf{c}$ *is preferred to* $\mathbf{f}$*. Therefore* $\mathbf{acg}$ *should be preferred to* $\mathbf{bfg}$*."*

We denote:

$$[(\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{f})] \xrightarrow{rt} (\mathbf{aceg}, \mathbf{befg})$$

[2]This amounts to allowing the tr scheme to operate on the quotient set of congruent classes of comparative statements, instead of representative of those classes.

We can note that Definition 9 is inefficient, as the search space of sequences of comparisons of length k is finite, but intractably large. The lack of specification can be overcome by reconstructing the missing intermediate steps, solving iteratively k fourth congruent problems (see Lemma 1), which can be done in $O(km^2)$ time.

Moreover, the rt scheme fulfills its role by allowing one-step derivation of proofs combining tr and cp derivations.

PROPOSITION 1. *If the premises $\mathcal{P} = \bigcup_{j=1}^{n}(A_j, B_j)$ allow to prove the conclusion (A, B) via the cp and tr schemes, then there is a list $[P'_1, \ldots, P'_k]$ of comparative statements of $\mathcal{P}$ such that $[P'_1, \ldots, P'_k] \xrightarrow{rt} (A, B)$.*

PROOF. By construction, the rt schemes subsumes the tr and cp schemes (so $\vdash_{tr,cp} \subset \vdash_{rt}$), and particularizes a proof combining tr and cp derivations (so $\vdash_{tr,cp} \supset \vdash_{rt}$). Moreover, chaining rt derivations is useless, because if $L_1 \xrightarrow{rt} (A, B)$ and $L_2 \xrightarrow{rt} (B, C)$, then $L_1 \& L_2 \xrightarrow{rt} (A, C)$, where & is list concatenation. Therefore, $\vdash_{rt} = \xrightarrow{rt}$. □

3.5 The decomposition scheme

Introduced in [2] and implementing cancellation properties of higher order, the decomposition scheme aims at leveraging the assumed additive property of the preference relation[3]. When preference is additive, preference statements translate into linear comparisons, that can be summed up. Then, the scores of items appearing on both sides cancel out, sometimes allowing to derive new comparisons.

Definition 10 (decomposition scheme (dec))**.** A premise $[(A_1, B_1), \ldots, (A_k, B_k)]$ and a conclusion (A, B) satisfy the *decomposition scheme* when each comparative statement (A_i, B_i) is disjoint and, for all items $j \in A \setminus B$, there are as many occurrences of j in the sets $A_1, \ldots, A_k$ as there are in the sets $B_1, \ldots, B_k$ plus one; for all items $j \in B \setminus A$, there are as many occurrences of j in the sets $B_1, \ldots, B_k$ as there are in the sets $A_1, \ldots, A_k$ plus one; and for any item j neither in A nor B, or both in A and B, there are as many occurrences of j in the sets $A_1, \ldots, A_k$ as there are in the sets $B_1, \ldots, B_k$, i.e. $\forall\, j \in [m]$

$$\sum_{i=1}^{k} |A_i \cap \{j\}| + |B \cap \{j\}| = \sum_{i=1}^{k} |B_i \cap \{j\}| + |A \cap \{j\}|$$

PROPOSITION 2. *If the preference $\gtrsim$ is additive, then the decomposition scheme is correct w.r.t. $\gtrsim$.*

EXAMPLE 6. *Consider the following decomposition scheme:*

$$[(\mathbf{bc}, \mathbf{de}), (\mathbf{efg}, \mathbf{ac})] \xrightarrow{dec} (\mathbf{bfg}, \mathbf{ad})$$

Assuming that the preference $\gtrsim$ is additive, and that both $\mathbf{bc} \gtrsim \mathbf{de}$ and $\mathbf{efg} \gtrsim \mathbf{ac}$. From the first comparison, we deduce that $\omega_{\mathbf{b}} + \omega_{\mathbf{c}} \geq \omega_{\mathbf{d}} + \omega_{\mathbf{e}}$; from the second that $\omega_{\mathbf{e}} + \omega_{\mathbf{f}} + \omega_{\mathbf{g}} \geq \omega_{\mathbf{a}} + \omega_{\mathbf{c}}$. By summation, we derive $\omega_{\mathbf{e}} + \omega_{\mathbf{f}} + \omega_{\mathbf{g}} + \omega_{\mathbf{b}} + \omega_{\mathbf{c}} \geq \omega_{\mathbf{d}} + \omega_{\mathbf{e}} + \omega_{\mathbf{a}} + \omega_{\mathbf{c}}$.

Then, as it is illustrated by Fig. 2 by cancelling $\omega_{\mathbf{e}}$ and $\omega_{\mathbf{c}}$ on both sides (this is actually an instance of second order cancellation*, because it is performed across two comparative statements), we obtain $\omega_{\mathbf{f}} + \omega_{\mathbf{g}} + \omega_{\mathbf{b}} \geq \omega_{\mathbf{d}} + \omega_{\mathbf{a}}$, hence $\mathbf{bfg} \succ_{\omega} \mathbf{ad}$.*

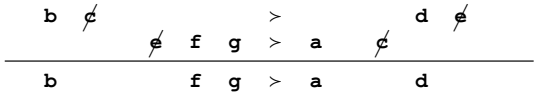


Figure 2: Decomposition scheme: graphical representation

The decomposition scheme strictly generalizes the schemes introduced previously.

PROPOSITION 3. *If a premise $[(A_1, B_1), \ldots, (A_k, B_k)]$ and a conclusion (A, B) satisfy the transitive or reduced transitive scheme, then they satisfy the decomposition scheme.*

We note that the decomposition scheme is commutative by construction. Nevertheless, it does not seem very satisfying as an explanation device, as the cancellation properties of higher order it enacts are complex and of low normative appeal–actually, even though they are syntactically transductive, their justification derives from the additive form we strive to circumvent.

In general, instances of the decomposition scheme do not satisfy the reduced transitive scheme. The sequence of comparative statements of the premise can not be interpreted as *ceteris paribus* justifications of comparative statements between states, because at some point, they either require to add an item to a state where it is already present, or to remove an item from a state where it is absent.

EXAMPLE 7. *The premise $[(\mathbf{bc}, \mathbf{de}), (\mathbf{efg}, \mathbf{ac})]$ and the conclusion $(\mathbf{bfg}, \mathbf{ad})$ satisfy the decomposition scheme, but they cannot be interpreted as a sequence of comparative statements because e.g. the initial state of the conclusion $\mathbf{bfg}$ does not contain $\mathbf{c}$.*

This situation can be avoided when the states mentioned in the premise are all pairwise disjoint.

PROPOSITION 4. *Let $[(A_1, B_1), \ldots, (A_k, B_k)]$ a premise and (A, B) a conclusion satisfying the decomposition scheme. Every permutation σ of the indices $[k]$ makes the premise $[(A_{\sigma(1)}, B_{\sigma(1)}), \ldots, (A_{\sigma(k)}, B_{\sigma(k)})]$ and conclusion (A, B) satisfy the reduced transitive scheme if, and only if, the states $A_1, \ldots, A_k, B_1, \ldots, B_k$ are pairwise disjoint.*

PROOF. (Sketch) ($\Leftarrow$) When all the states $A_1, \ldots, A_k, B_1, \ldots, B_k$ are pairwise disjoint, each item of $A \setminus B$ appears exactly once in the B_i and never in the A_i, each item of $B \setminus A$ appears exactly once in the A_i and never in the B_i, and items in both sets or neither never appear (incidentally making the transaction III (see Definition 2)). Consequently, items of the set $\bigcup_{i=1}^{k} A_i$ can be removed in any order from A and those in the set $\bigcup_{i=1}^{k} B_i$ can be added in any order, so as to accrue to B. □

This motivates the definition of the *covering scheme*.

[3]This decomposition scheme is less general than the so-called *syntactic cancellative* scheme described in [2], as it does not allow for repetition of the conclusion. This has been shown to reduce expressiveness.

3.6 The covering scheme

In this scheme a list of comparative statements $[(A_1, B_1), \ldots, (A_k, B_k)]$ supports a conclusion (A, B) if, and only if, the *pros* $A_1, \ldots, A_k$ partition $A \setminus B$ and the *cons* $B_1, \ldots, B_k$ partition $B \setminus A$.

Definition 11 (covering scheme (*cov*))**.** An instance of the *covering scheme* is an instance of the decomposing scheme where all the states $A_1, \ldots, A_k, B_1, \ldots, B_k$ are pairwise disjoint.

The covering scheme is commutative and independent of irrelevant items by construction. It particularizes the reduced transitive schemes – while circumventing the tedious scheduling of comparative statements – and, as a corollary, it is correct under much milder conditions than the decomposition scheme.

PROPOSITION 5. *If the preference $\succsim$ is transitive and satisfies cancellation, then the covering scheme is correct w.r.t. $\succsim$.*

The covering scheme describes exactly the *moral algebra* introduced by Benjamin Franklin (see [21]) to infer preferences.

EXAMPLE 8. *Consider the conclusion:* $(\mathbf{bfg}, \mathbf{cde})$. *The premise* $[(\mathbf{fg}, \mathbf{c}), (\mathbf{b}, \mathbf{de})]$ *constitute a covering scheme:*

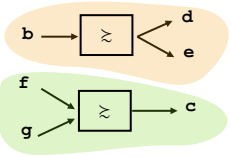
$$[(\mathbf{fg}, \mathbf{c}), (\mathbf{b}, \mathbf{de})] \xrightarrow{cov} (\mathbf{bfg}, \mathbf{cde})$$

The covering scheme particularizes both the reduced transitive scheme, and the decomposition scheme. As such, it gives us the best of both world, in a sense. On the one hand, it formalizes a proof, articulating transitive and *ceteris paribus* derivations, that can be presented to the explainee as a diagram, such as in Fig. 3a, or narratively such as in Fig. 3c (for hotel comparisons for instance). On the other hand, the premises can be understood as grouping some cons with some stronger pros, so as to "cover" the cons, and can be presented visually to the explainee such as in Fig. 3b.

(a) Covering Scheme: proof diagram of Ex. 8

$$\left.\begin{array}{l} \mathtt{fg} \succ \mathtt{c} \xrightarrow{cp} \mathtt{bfg} \succ \mathtt{bc} \\ \mathtt{b} \succ \mathtt{de} \xrightarrow{cp} \mathtt{bc} \succ \mathtt{cde} \end{array}\right\} \xrightarrow{tr} \mathtt{bfg} \succ \mathtt{cde}$$

(b) Covering Scheme: a visual representation of Ex.8



(c) Covering Scheme: a narrative representation of Ex.8

"As, all other things being equal, having free breakfast and wifi access is preferred to having a swimming pool ($\mathbf{fg}, \mathbf{c}$*), and being close to the city is preferred than having a sports hall and a low tourist tax (*$\mathbf{b}, \mathbf{de}$*), we get that* $(\mathbf{bfg}, \mathbf{cde})$*"*

Figure 3: Three representations of the Covering Scheme

Note also that a single covering scheme of length k can be interpreted as $k!$ transitive schemes, as the validity of its premises does not depend on their order. Hence, for instance for our Ex. 8, two transitive schemes would correspond to this covering scheme: $[(\mathbf{bfg}, \mathbf{cdfg}), (\mathbf{cdfg}, \mathbf{cde})]$ or $[(\mathbf{bfg}, \mathbf{be}), (\mathbf{be}, \mathbf{cde})]$

We presented in this section different types of schemes representing different way of reasoning over preferences. We believe that these schemes may correspond to alternative explanation strategies.

One can note that it exist a logical dependency between the premises and conclusions satisfying the various schemes. Indeed, all instances satisfying the *cov* scheme satisfy the *rt* scheme, and the ones satisfying the latter satisfy the *dec* scheme (the converse is not true). This has an obvious consequence on the explainabilty relations: more general implies more explicative. Moreover, the conclusions of the *tr* scheme can all be obtained via the *rt* scheme, and those of the *cp* via the *dec*, *rt* or *cov* schemes.

4 EXPLAINING WITH SCHEMES

Section 3 was about engineering a deductive tool allowing to derive complex preferences from simpler one, in a *correct* manner with respect to the latent preference model. Given an *explanans*–a comparative statement that needs to be explained–it allows us to cast the *explanation problem* as an *abduction* problem of finding premises that satisfy some minimality requirement given a conclusion and a set of rules. We now investigate the relative expressiveness and computational complexity of explaining with the *rt*–see Def. 9–and *cov*–see Def. 11–schemes, together with the influence of the choice of the atomically simple statements.

From now on, we denote as $\mathcal{E}(s, \mathcal{A}_{\succsim}, k)$ the set of pairs (A, B) that can be derived using the scheme s from statements respecting the syntactic constraints of $\mathcal{A}$, the semantic constraint of being coherent with $\succsim$, and involving at most k premises. For a given pair (A, B) the explanation existence problem asks whether $(A, B) \in \mathcal{E}(s, \mathcal{A}_{\succsim}, k)$. By convention, this pair (A, B) is considered not self-evident.

4.1 Solving Explanation Problems with Schemes

As we have seen in the previous section, when preference is additive, all our reasoning schemes in {transitive, ceteris paribus, reduced transitive, decomposition, covering} are correct. However, we cannot expect them to be *complete*, even without any syntactic restriction on $\mathcal{A}$: when states $A \succ B$ are adjacent in the preference relation $\succsim$, i.e. when there is no other state X s.t. $A \succ X \succ B$, (A, B) are said to be a *critical pair*, see [9]. But this means in turn that the conclusion (A, B) cannot be obtained with the *rt* scheme–even less so by the *cov* and the *tr* schemes. These critical pairs are thus not explainable with these schemes (see Ex.9).

EXAMPLE 9. *The conclusion* $(\mathbf{bcd}, \mathbf{aefg})$ *is a critical pair. Indeed,* $\omega_{\mathbf{b}} + \omega_{\mathbf{c}} + \omega_{\mathbf{d}} = 262$, $\omega_{\mathbf{a}} + \omega_{\mathbf{e}} + \omega_{\mathbf{f}} + \omega_{\mathbf{g}} = 258$, *and it is not possible to exhibit another state from* $2^{[m]}$ *with a score* $\in]258, 262[$.

On the other hand, as soon as a pair is *not* critical, and provided that $\mathcal{A}$ does not put any syntactic constraint on the statements used, there must exist at least one explanation with schemes *tr* and *rt*. This means that the complexity of deciding the existence of an explanation for these schemes is directly related to that of deciding whether a pair is critical. We show that this problem is difficult.

THEOREM 1. *Given $\omega \in \mathbb{N}_0^m$, and $A, B \in 2^{[m]}$ such that $A \gtrsim B$ where $\gtrsim$ the additive preference relation induced from ω. Deciding whether (A, B) is a critical pair is Co-NP-complete.*

PROOF. Reduction from SUBSET-SUM [10]. In SUBSET-SUM we are given a set A of size m a positive size $w(a)$, for each $a \in A$, and a positive integer K. We ask whether there is a subset $A' \subseteq A$ such that the sum of the weights is exactly K. This problem is known to be NP-complete. We construct an instance of the critical pair problem as follows. We take a set C of $m+2$ criteria: for each element $a_i \in A$, we take a criteria c_i, of weight $2s(a_i)$. We add two other criteria: a_{n+1} of weight $2K-1$, and a_{n+2}, of weight $2K+1$. We ask whether the pair (X, Y) is critical, where $X = \langle 0, \ldots, 1, 0\rangle$ and $Y = \langle 0, \ldots 1\rangle$. Note that X and Y have respective weights of $2K-1$ and $2K+1$, hence we are looking for an intermediate alternative of weight exactly $2K$, an even number. We claim that the answer to this question is no iff the original SUBSET-SUM problem is a yes-instance. To see this, observe that all the weights in our critical pair instance are even, except that of c_{n+1} and c_{n+2} which are odd. As $s(c_{n+2}) > 2K$, it certainly cannot be part of the solution. Furthermore, the solution cannot include c_{n+1}, as in that case it would be the only odd weight, and then the sum would be odd. We are left with criteria $c_1, \ldots c_n$, whose weights are precisely twice the weight in the original SUBSET-SUM problem. □

As a corollary, it is difficult to decide whether an explanation exists with these schemes.

COROLLARY 1. *Given $\omega \in \mathbb{N}_0^m$ $A, B \in 2^{[m]}$ such that $A \gtrsim B$, where $\gtrsim$ is the additive preference relation induced from ω, and $\mathcal{A}$ the set of statements $\Delta(m, m)$. Deciding whether $(A, B) \in \mathcal{E}(s, \mathcal{A}_{\gtrsim}, +\infty)$ is NP-complete for $s \in \{rt, tr\}$.*

For our scheme of choice *cov*, we can prove intractability through an independent proof.

THEOREM 2. *Given $\omega \in \mathbb{N}_0^m$, $A, B \in 2^{[m]}$ such that $A \gtrsim B$ where $\gtrsim$ is the additive preference relation induced from ω and $\mathcal{A}$ the set of statements $\Delta(m, m)$. Deciding whether $(A, B) \in \mathcal{E}(cov, \mathcal{A}_{\gtrsim}, +\infty)$ is NP-complete.*

4.2 Explaining with Atomic Statements

We now turn our attention towards explanations which put syntactic restrictions on the sets of atomic elements used, namely $\Delta(1, 1)$, $\Delta(1, m)$, and $\Delta(m, 1)$ (see Sect.2).

THEOREM 3. *When $\gtrsim$ is additive, $\mathcal{E}(cov, \mathcal{A}_{\gtrsim}, \infty)$ is transitive when $\mathcal{A}_{\gtrsim} \in \{\Delta(1, 1), \Delta(1, m), \Delta(m, 1)\}$.*

PROOF. Suppose $[(A_1, B_1), \ldots, (A_k, B_k)] \xrightarrow{cov} (A, B)$ and $[(A'_1, B'_1), \ldots, (A'_k, B'_k)] \xrightarrow{cov} (B, C)$. We show the conclusion (A, C) is yielded by applying the covering scheme to some premise. It is easy to check that $[(A_1, B_1), \ldots, (A_k, B_k), (A'_1, B'_1), \ldots, (A'_k, B'_k)] \xrightarrow{dec} (A, C)$, and we only need to prove the sets $A_1, \ldots, A_k, A'_1, \ldots, A'_{k'}, B_1, \ldots, B_k, B'_1, \ldots, B'_{k'}$ are pairwise disjoint. We already know that the $\langle A_i, B_i\rangle$ and the $\langle A_{i'}, B_{i'}\rangle$ are pairwise disjoint. Moreover, the A_i are contained in $A \setminus B$ while the $A'_{i'}$ are in $B \setminus C$, so they do not intersect (same for the B_i and $B'_{i'}$). The only intersections left to consider are $A_i \cap B'_{i'}$ and $B_i \cap A'_{i'}$. Suppose w.l.o.g. $A_i \cap B'_{i'} \neq \emptyset$. Because of the syntactic constraints $\mathcal{A}$ we consider, this intersection is a singleton $\{j\}$. We delete the comparative statements (A_i, B_i) and $(A'_{i'}, B'_{i'})$ from the premise, and replace them with the comparative statement $(A_i \cup A'_{i'} \setminus \{j\}, B_i \cup B'_{i'} \setminus \{j\})$. This comparative statement belongs to $\mathcal{A}$, and also to $\gtrsim$ because it is additive (by summation of the inequalities characterizing $A_i \gtrsim B_i$ and $A'_{i'} \gtrsim B'_{i'}$ and cancellation of the terms ω_j appearing on both sides). By iterating this operation, we obtain a covering scheme supporting (A, C) of size no greater than $k + k'$. □

As a corollary, when restricting the atoms to mention solely one pro vs any number of cons (resp. any number of pros vs one con), the *rt* scheme is not more expressive than the *cov* one. This is not the case when we allow to mix these atoms (as illustrated by Ex. 10).

EXAMPLE 10. *The premise $[(\mathbf{b}, \mathbf{ge}), (\mathbf{fg}, \mathbf{c})]$ and the conclusion $(\mathbf{bf}, \mathbf{ce})$ satisfy the rt scheme (as it is illustrated in the figure below). However, this is not the case for the cov scheme as $\omega_f < \omega_c + \omega_e$ and $\omega_b < \omega_c + \omega_e$*

$$\left.\begin{array}{l}(\mathbf{b}, \mathbf{eg}) \xrightarrow{cp} (\mathbf{bf}, \mathbf{efg}) \\ (\mathbf{fg}, \mathbf{c}) \xrightarrow{cp} (\mathbf{efg}, \mathbf{ce})\end{array}\right\} \xrightarrow{tr} (\mathbf{bf}, \mathbf{ce})$$

PROPOSITION 6. *For any positive integer k, when $\mathcal{A} \in \{\Delta(1, 1), \Delta(1, m), \Delta(m, 1)\}$ and $\gtrsim$ is additive, $\mathcal{E}(rt, \mathcal{A}_{\gtrsim}, k) = \mathcal{E}(cov, \mathcal{A}_{\gtrsim}, k)$.*

One may wonder whether these atomic statements make the problem computationally simpler to handle. While the answer is known to be positive for $\mathcal{A} = \Delta(1, 1)$ [1], we show that from $k \geq 2$ the problem is difficult.

THEOREM 4. *Given $\omega \in \mathbb{N}_0^m$ $A, B \in 2^{[m]}$ such that $A \gtrsim B$ where $\gtrsim$ is the additive preference relation induced from ω, and $\mathcal{A} = \Delta(1, k)$. When $k \geq 2$, deciding whether $(A, B) \in \mathcal{E}(cov, \mathcal{A}_{\gtrsim}, +\infty)$ is NP-complete.*

When using the *cov* scheme, the length of explanations is bounded by the number m of items. There are thus a range of values of m for which finding an explanation might prove too difficult for a human, but can be easily achieved by a machine, either with a solver or even by brute force. This is a plea for the use of artificial explainers.

5 EMPIRICAL COMPLETENESS OF THE COVERING SCHEME

The results of Section 4 establish that the sets of atomic statements $\Delta(1, m)$ or $\Delta(m, 1)$ using the *cov* scheme discharge us from the task of sequencing the explanations. Of course, this comes at a price, as some pairs which may be explained otherwise may not be explained with these statements. This section provides insights regarding the "empirical completeness" of the *cov* scheme with such statements.

In general, given an additive preference relation, noted $\mathcal{R}$, the set $\mathcal{A}_{\mathcal{R}}$ of comparative statements that can be used as a premise for explaining the conclusion (A, B) such that $(A, B) \in \mathcal{R}$ and $(A, B) \notin \mathcal{A}_{\mathcal{R}}$ is the following:

$$\mathcal{A}_{\mathcal{R}} = [\Delta(1, m) \cup \Delta(m, 1)] \cap \mathcal{R}$$

In our context, we consider preference relations over states that are representable by *additive linear order on the algebra of subsets of a finite set* (see [9]). We assume the following ordering over the singleton states: $\mathbf{a} \succ \mathbf{b} \succ \mathbf{c} \succ \mathbf{d} \succ \mathbf{e} \succ \mathbf{f} \succ \ldots$. Thus, we denote by $T^m_\succ = \{(A, B) \in 2^{[m]} \times 2^{[m]} : A \succ B \text{ and } A \cap B = \emptyset\}$, and we have $\mathcal{A}_\succ \subseteq T^m_\succ$.

The number of additive linear orders on $2^{[m]}$ grows very quickly [9, 15]. It corresponds to 14 for $m = 4$, but for $m = 7$ we already have more than 200 million orders. Technically, we have been able to generate all additive linear orders for $m \in [\![4; 6]\!]$. Thus, to decide whether $(A, B) \in \mathcal{E}(cov, \mathcal{A}_\succ, \infty)$, we used a Mixed Integer Linear Program solver. Finally, to evaluate the proportions of pairs $T^m_\succ \backslash \mathcal{A}_\succ$ which are explainable, we compute for each *additive linear order* $\succ$ given m, the following value

$$\mathfrak{M}_{m,\succ} = \frac{|\mathcal{E}(cov, \mathcal{A}_\succ, \infty) \cap T^m_\succ \backslash \mathcal{A}_\succ|}{|T^m_\succ \backslash \mathcal{A}_\succ|}$$

m	Minimum	Median	Maximum	$\lvert T^m_\succ \backslash \mathcal{A}_\succ \rvert$
4	66.7%	66.7%	100%	3
5	72.0%	80.0%	100%	25
6	78.46%	84.62%	100%	130

Table 1: $\mathfrak{M}_{m,\succ}$ for $m \in [\![4; 6]\!]$ $\quad \forall \succ$

Table 1 summarizes the minimum, median and maximum values obtained over the, respectively, 14, 516, and 124187 additive linear orders (for $m = 4, 5, 6$). We notice that both the minimum and the median values of $\mathfrak{M}_{m,\succ}$ increases with m. Regarding the maximum values, we note that there are all equal to 100% which means that for all $m \in [\![4; 6]\!]$, there exists at least one additive linear order $\succ$ for which all pairs of states are explainable.

Looking more globally at the set of values in the Table 1, we can say that a significant majority of the pairs of $T^m_\succ \backslash \mathcal{A}_\succ$ are explainable. For example, for $m = 6$, more than 3 pairs out of 4 are explainable regardless of the additive linear order considered.

Of course, the explainability of an arbitrary couple (A, B) depends of its characteristics w.r.t the ranking of the criteria which compose them in the ordering over the singleton states. For example, couples as (**ac**, **bd**) will always be explainable since $\mathbf{a} \succ \mathbf{b}$ and $\mathbf{c} \succ \mathbf{d}$. However, it will be more difficult to decide for pairs like (**ad**, **bc**) or (**ae**, **bcd**) or (**bde**, **acf**).

6 RELATED WORK AND CONCLUSION

Recently, [16] have explored explanations in the context of decision of linear classifiers. They focus on PI-explanations (or sufficient reasons), that is, explanations providing reasons sufficient to explain a given decision, regardless of the value of other criteria [7, 8], a strategy of explanation different from ours as mentioned in the introduction. Our view of explanations as cognitively bounded deductive proofs is reminiscent of the *bounded proof systems* proposed in the context of description logic [12]. Also, a similar step-wise approach has been studied in the context of constraint satisfaction problems [3]. Finally, explanations based on axioms has been advocated in computational social choice [5, 19]. In particular, the recent work of [4] also exploits axioms studied in voting theory to produce explanations for collective decisions, but applied to a different setting (voting), and using different proof techniques (tableau methods).

We propose a framework to explain comparisons stemming from an additive model. The framework comes with different schemes for reasoning with preferences, and we focus on a specific one: the covering scheme. Moreover, for cognitive purposes, we propose to restrict explanations to sets of atomic elements which prefer a pro over a group of cons, or a group of pros over a single con. The covering-based explanation engine with restricted sets of atomic elements is not complete, but empirical investigations show that explanations can be computed in a large proportion of cases.

Moreover, providing an argument scheme along with the result of a comparative statement opens the possibility to discuss or challenge this result. This is made possible through what is called critical questions [20], a tool associated with argument schemes representing attacks or criticisms that, if not answered adequately, falsify the argument fitting the scheme. This naturally leads to the long-term perspective of the interactive nature of the explanation process: these schemes should be integrated into a dialectical process, whereby the end-user should be able to contest [18], while on the other hand the system should gain knowledge about the preferences of the user through an indirect elicitation process [14]. Smoothly interleaving explanation and recommendation calls for designing mixed initiative systems [13] where the user may be active in challenging the system, and the system adaptive in its responses.

REFERENCES

[1] Kh. Belahcene, Ch. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. 2017. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision* 82, 2 (2017), 151–183.

[2] Kh. Belahcene, Ch. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. 2019. Comparing options with argument schemes powered by cancellation. In *Proc. 28th IJCAI*. AAAI Press, 1537–1543.

[3] B. Bogaerts, E. Gamba, and T. Guns. 2021. A framework for step-wise explaining how to solve constraint satisfaction problems. *Artif. Intell.* 300 (2021), 103–550.

[4] A. Boixel, U. Endriss, and R. Haan. 2022. A Calculus for Computing Structured Justifications for Election Outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4859–4866.

[5] O. Cailloux and U. Endriss. 2016. Arguing about Voting Rules. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. ACM, 287–295.

[6] S. Coste-Marquis and P. Marquis. 2020. From Explanations to Intelligible Explanations. In *1st International Workshop on Explainable Logic-Based Knowledge Representation (XLoKR'20)*. Workshop at KR'20.

[7] A. Darwiche and A. Hirth. 2020. On The Reasons Behind Decisions. In *Proceedings of the 24th ECAI*.

[8] A. Darwiche and P. Marquis. 2021. On Quantifying Literals in Boolean Logic and its Applications to Explainable AI. *Journal of Artificial Intelligence Research* 72 (2021), 285–328.

[9] P. C. Fishburn, A. Pekec, and J. A. Reeds. 2002. subset comparisons for additive linear orders. *Mathematics of Operations Research* 27 (2002), 227–243.

[10] M. R. Garey and D. S. Johnson. 1979. *Computers and Intractability, a Guide to the Theory of NP-completeness*. Freeman.

[11] J Hammond, R. Keeney, and H Raiffa. 1998. Even Swaps: A Rational Method for Making Trade-offs. *Harvard business review* 76 (03 1998), 137–8, 143.

[12] M. Horridge, S. Bail, B. Parsia, and U. Sattler. 2013. Toward Cognitive Support for OWL Justifications. *Know.-Based Syst.* 53 (nov 2013), 66–79.

[13] E. Horvitz. 2000. Uncertainty, Action, and Interaction: In Pursuit of Mixed-Initiative Computing. *Intelligent Systems* (2000), 17–20.

[14] Ch. Labreuche, N. Maudet, W. Ouerdane, and S. Parsons. 2015. A Dialogue Game for Recommendation with Adaptive Preference Models. In *Proceedings of the 14th AAMAS*. 959–967.

[15] D. Maclagan. 1998. Boolean Term Orders and the Root System B_n. *Order* 15 (1998), 279–295.

[16] J. Marques-Silva, Th. Gerspacher, M. Cooper, A. Ignatiev, and N. Narodytska. 2020. Explaining Naive Bayes and Other Linear Classifiers with Polynomial

Time and Delay. In *Advances in Neural Information Processing Systems*, Vol. 33. 20590–20600.
[17] T. Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.
[18] D. K. Mulligan, D. Kluttz, and N. Kohli. 2020. Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. In *After the Digital Tornado*. Cambridge University Press.
[19] A. D. Procaccia. 2019. Axioms Should Explain Solutions. *The Future of Economic Design* (2019).
[20] D. Walton. 1996. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum.
[21] William B. Willcox (Ed.). 1975. *The Papers of Benjamin Franklin*. New Haven and London, Yale University Press, 299–300.