Des explications par étapes pour le modèle additif

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

Des explications par étapes pour le modèle additif

Manuel Amoussou[1] **Khaled Belahcene**[2] **Nicolas Maudet**[3] **Vincent Mousseau**[1] **Wassila Ouerdane**[1]

[1]MICS, CentraleSupélec, Université Paris-Saclay, France
[2]Heudiasyc, Université de Technologie de Compiègne, CNRS, France
[3]Lip6, Sorbonne Université, CNRS, France

{manuel.amoussou, vincent.mousseau, wassila.ouerdane}@centralesupelec.fr,
khaled.belahcene@hds.utc.fr, nicolas.maudet@lip6.fr

Résumé

Nous nous intéressons au problème du calcul d'explications de comparaisons par paires issues d'un modèle de préférence additif. La structure de ces explications s'appuie sur une décomposition de la paires à expliquer sous la forme d'une séquence de jugements de préférence. Pour que cette explication soit comprise et acceptée par celui à qui elle est destinée, chaque élément de la séquence construite doit être aussi intelligible et cognitivement simple que possible. Ainsi, nous proposons plusieurs schémas d'arguments permettant d'inférer de nouvelles connaissances, sous forme de comparaisons par paires, à partir de celles précédemment validées. Ces schémas d'arguments dont nous garantissons la correction, exploitent un certain nombre de propriétés du modèle additif. Nous préconisons spécifiquement l'utilisation du schéma couverture car il répond à certaines propriétés souhaitables pour les explications. Imposer que la décomposition soit faite en éléments cognitivement simples se fait au prix de l'exhaustivité. Cependant, les résultats expérimentaux montrent que nous sommes capables de fournir des explications dans une large proportion de cas.

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 explanatory 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.

1 Introduction

Dans cet article, nous abordons le problème de la production d'explications progressives pour les comparaisons par pairess d'alternatives basées sur des modèles de décision bien établis. Les alternatives sont caractérisées par un certain nombre de critères. L'idée principale est de décomposer la recommandation en affirmations simples présentées au destinataire de l'explication. L'ensemble de la séquence de comparaisons doit soutenir formellement la recommandation. Les explications que nous visons sont donc *contrastives*, dans le sens où la décision à expliquer compare deux alternatives, et *exactes* (par opposition à *heuristiques*) dans le sens où nous fournissons des garanties que l'explication produite est correcte par rapport au modèle sous-jacent. Il est également courant de distinguer entre les explications *locales* (lorsqu'elles se concentrent sur une recommandation spécifique) et les explications *globales* (lorsqu'elles traitent du modèle en général) : notre approche est globalement fidèle au modèle, et localement pertinente pour la comparaison par paires à expliquer. Enfin, même si cet aspect n'est pas détaillé dans ce travail, la perspective est de donner à celui qui re**c**coit l'explication la possibilité d'accepter ou de contredire ces affirmations. Pour rendre les choses concrètes, nous commen**c**cons par l'exemple suivant.

Exemple et motivations Nous considérons sept critères ($\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g}$), chacun décrit sur une échelle à

deux niveaux, ce qui facilite la représentation symbolique des alternatives (*e.g* des hôtels). Chaque alternative peut être représentée par son vecteur d'évaluation ($s_1 = (\text{✗}, \text{✗}, \text{✓}, \text{✓}, \text{✓}, \text{✓})$) ou plus succinctement par le sous-ensemble de critères sur lesquels elle est évaluée positivement ($s_1 = \{\mathbf{cdefg}\}$). De plus, pour chaque critère, la valeur symbolisée par ✓ est plus désirable que la valeur symbolisée par ✗ (*e.g* un hôtel avec un le petit déjeuner inclus est mieux que sans).

TABLE 1 – Évaluation de deux alternatives

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

L'agrégation des critères se fait à l'aide d'une fonction de score additif, attribuant des poids aux différents critères. Cette fonction est comme suit :

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

Par exemple, le score de s_1 correspond à $score(s_1) = 77 + 59 + 52 + 41 + 37 = 276$ tant dis que celui de s_2 is : $score(s_2) = 128+59 = 187$. Il est également utile d'encoder la comparaison de deux alternatives sous la forme d'un vecteur $\{-1, 0, +1\}^n$ d'arguments en faveur (PRO) ou contre (CON) s_1, ou neutres (NEU). Dans notre exemple, PRO = $\{\mathbf{c}, \mathbf{e}, \mathbf{f}, \mathbf{g}\}$, CON = $\{\mathbf{a}\}$, tandis que NEU = $\{\mathbf{b}, \mathbf{d}\}$:

Les explications peuvent prendre plusieurs formes. Nous listons différentes explications possibles au fait que s_1 est préféré à s_2 :

(i) La première approche que nous nommerons par (*divulgation du modèle*) consisterait à fournir le calcul complet du score pour les deux options, comme illustré ci-dessus. Mais en remarquant que **d**, un argument neutre, satisfait à la fois par s_1 et s_2, nous pourrions l'omettre et fournir simplement la somme des arguments PRO contre les arguments CON.

(ii) l'approche *contre-factuelle* recherche la modification minimale de l'entrée qui changerait le résultat. Par exemple, nous pourrions affirmer que, si s_2 avait satisfait **b**, s_2 aurait été recommandé plutôt que s_1. Ou (en affectant l'autre alternative cette fois), si s_1 n'avait pas satisfait **cd**.

(iii) En suivant une approche *prime implicant*, nous pourrions produire les arguments suffisants pour expliquer la décision. Dans notre cas, deux explications possibles pourraient être données : (1) étant donné que **bd** sont des arguments neutres, les arguments PRO **cef** sont suffisants pour contrer tout ensemble de CON arguments. En particulier, cela montre que la décision resterait la même même si **g** était un argument CON. Et (2) étant donné que **b** est un argument neutre, les arguments PRO **cefg** sont suffisants pour contrer tout ensemble d'arguments CON. En particulier, cela montre que la décision resterait la même même si **d** était un argument CON.

(iv) En suivant une approche *pas à pas ou progressive*, nous pourrions exposer une collection d'affirmations visant à prouver la décision. Par exemple, nous pourrions affirmer que **cdefg** est préféré à **ac**, et que **ac** est préféré à **ad**, de sorte que notre conclusion devrait tenir, suivant un raisonnement transitif. Ou, en utilisant une logique différente, nous pourrions affirmer que **cd** est préféré à **a**, tandis que **efg** est préféré à **d**, ce qui justifie entièrement notre décision.

Cet exemple nous permet d'illustrer certains principes clés de l'explication (voir par *e.g.* [18, 6]). :

— *L'intelligibilité du langage*— nous souhaitons que les explications soient transmises dans un langage qui soit significatif pour la personne qui re**c**coit cette explication. Dans notre exemple, les poids des critères utilisés dans les calculs peuvent ne pas être facilement interprétés par le destinataire de l'explication [1]

— *pertinence*— nous souhaitons que les explications se concentrent sur les informations pertinentes. Dans notre exemple, comme nous l'avons remarqué, la mention d'arguments neutres peut sembler non pertinente et doit être évitée si possible.

— *simplicité cognitive*— nous souhaitons que les explications soient "faciles à traiter" par le destinataire de l'explication. Ceci peut être instancié de différentes manières : les explications de type "prime implicante" cherchent des (sous-ensemble) raisons suffisantes minimales, tandis que les explications pas à pas (progressives) font appel à des comparaisons intermédiaires impliquant un nombre limité de critères.

Notre ambition dans cet article est de développer un modèle d'explications pas à pas (progressif) basé sur des principes et limité sur le plan cognitif. Comme l'illustre notre exemple, il peut y avoir différentes "logiques" en jeu lors de la combinaison des affirmations. Pour en tenir compte, nous décrivons un certain nombre de *schémas* pour de telles explications dans le contexte d'une comparaison basée sur un modèle de somme pondérée (Section 3). Par approche basée sur des principes, nous entendons que chaque schéma est attaché à un certain nombre de propriétés bien comprises du modèle de décision sous-jacent, que nous rendons explicites et discutons dans cet article. Le calcul résultant est correct (Section 4). Par "cognitivement limité", nous entendons que nos affirmations seront contraintes de manière à rester faciles à comprendre par le destinataire de l'explication. Cela a pour conséquence de rendre le calcul résultant *non* complet, mais nous explorons cette question en détail et fournissons plusieurs éléments montrant que notre ap-

1. Un autre aspect, non étudié ici, est qu'il peut ne pas être adéquat de divulguer entièrement le modèle pour des questions de confidentialité ou de manipulation.

proche est satisfaisante en termes de complétude empirique (Section 5).

2 Notre modèle

Nous considérons un ensemble d'éléments ou d'items, noté $[m]$. Une *comparaison par paires* est une paires d'alternatives $(A, B) \in 2^{[m]} \times 2^{[m]}$, interprétée comme une déclaration de préférence 'A est préféré à B'.

Les schémas Notre objectif est de fournir un langage formel et un mécanisme de raisonnement permettant de soutenir (expliquer) de tels comparaisons par pairess. Nous nous appuyons sur la notion de *schéma d'arguments*, c'est-à-dire un opérateur liant une séquence de comparaisons, appelée prémisse, satisfaisant certaines conditions, à une autre comparaison appelé conclusion [21]. Comme nous traitons des préférences, les schémas d'arguments sont des moyens de dériver de nouvelles préférences à partir de celles qui ont été établies précédemment. Il est à noter que tous nos schémas opèrent sur le même ensemble de prémisses – des séquences finies de comparaisons par pairess, représentés par des listes entre crochets – et le même ensemble de conclusions – de comparaisons par pairess dans des $2^{[m]} \times 2^{[m]}$. Nous désignerons un schéma arbitraire s comme suit :

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

Correction Le fait que nos schémas d'arguments nous permettent uniquement de dériver des conclusions cohérentes avec la relation de préférence est capturé par la notion de correction :

Définition 1 *Un schéma d'argument est* correct *par rapport à une relation de préférence $\succsim$ si, lorsque toutes les prémisses appartiennent à $\succsim$, alors la conclusion appartient également à $\succsim$.*

A ce stade, nous laissons la relation de préférence non spécifiée, mais dans la section 3 nous approfondirons ce lien entre les propriétés des relations de préférence et les schémas.

Des affirmations simples L'approche *par étapes* est formalisée par l'exigence de *simplification* par rapport à la difficulté relative d'un énoncé.

Définition 2 *Une paires composée d'une prémisse $[(A_1, B_1), \ldots (A_k, B_k)]$ et d'une conclusion (A, B) est* simplifiante *lorsque la prémisse est moins difficile que la conclusion.*

Nous pensons que cette définition est très générale, car elle capture l'un des objectifs de l'explication. Pour être applicable, cependant, elle nécessite de spécifier la difficulté relative d'une prémisse et d'une conclusion.

Nous introduisons un modèle spécifique permettant de dériver la difficulté relative des énoncés, où cette difficulté est purement syntaxique et résulte directement du nombre d'items impliqués dans la comparaison par paires.

Définition 3 (Difficultés des affirmations) *La* difficulté *d'une comparaison par paires $(A, B) \in 2^{[m]} \times 2^{[m]}$ est la paires ordonnée d'entiers $(|A|, |B|)$. Par conséquent, on dit qu'une comparaison (A, B) est* moins difficile que *qu'une autre comparaison (A', B') lorsque $|A| \leq |A'|$, $|B| \leq |B'|$ et au moins une comparaison est stricte. Une séquence de comparaisons par paires $[(A_1, B_1), \ldots (A_k, B_k)]$ est* moins difficile qu'une *comparaison (A, B) lorsque toutes les comparaisons (A_i, B_i) sont moins difficiles que (A, B). Enfin, nous définissons des* classes de difficulté *de comparaisons par paires en fixant des limites supérieures à la difficulté : pour tous les entiers p, q de 0 à m, soit $\Delta(p, q) = \{(A, B) \in 2^{[m]} \times 2^{[m]} : |A| \leq p, |B| \leq q\}$.*

Notons $\mathcal{A}$ l'ensemble des éléments syntaxiquement atomiques, ceux qui sont considérés comme évidents et légitimes pour être utilisés comme étapes d'une explication pour le destinataire de l'explication considéré. Nous utiliserons les classes de difficulté $\Delta(p, q)$ pour spécifier cet ensemble. Dans le contexte de l'explication des préférences entre un sous-ensemble d'items désirés, certaines valeurs de la paires (p, q) présentent un intérêt particulier : les $\Delta(m, m)$ sont des énoncés non restreints ; les comparaisons dans $\Delta(m, 0)$ représentent des énoncés de dominance Pareto ; les comparaisons dans $\Delta(1, 1)$ peuvent être interprétées comme des *swaps* [12], représentant l'échange d'un critère contre un autre ; ceux en $\Delta(1, m)$ ou en $\Delta(m, 1)$ représentent un seul élément plus fort ou plus faible qu'un sous-ensemble d'autres, respectivement considérés comme un argument pour ou contre.

Par exemple, dans le contexte de comparaisons d'hôtels, un argument dans $\Delta(1, 1)$ pourrait être *"nous préférons avoir un petit-déjeuner gratuit qu'un accès wifi gratuit"*. Un argument dans $\Delta(1, 2)$ pourrait être *"Nous préférons avoir une piscine que le petit-déjeuner et le wifi gratuits"*. Pour comprendre à quel point il peut être difficile d'interpréter des arguments d'ordre supérieur, considérons des arguments dans $\Delta(2, 2)$. Ceux-ci pourraient correspondre à *"Un petit-déjeuner gratuit et un accès wifi sont préférables à avoir une piscine et à être proche du centre ville"*.

Dans la section 5, nous étudierons comment le fait de limiter l'explication à l'utilisation de ces classes d'énoncés simples affecte la capacité à produire des explications.

Explication basée sur des schémas. Les énoncés atomiques évidents limitent la difficulté de chaque étape d'une explication. Comme une explication est une séquence de

tels énoncés, nous cherchons également à produire des explications correctes de longueur minimale :

Définition 4 (Le problème d'explication) *Étant donné une comparaison par paires* $(A, B) \in 2^{[m]} \times 2^{[m]}$, *une relation de préférence* $\succsim$, *un ensemble d'énoncés* $\mathcal{A}$ *appartenant à* $\succsim$, *un ensemble de schémas* $\mathcal{S}$, *et un entier positif* k : *existe-t-il un entier positif* $k' \leq k$, *une liste de longueur* k' *d'énoncés* $[(A_1, B_1), \ldots, (A_{k'}, B_{k'})]$ *appartenant tous à* $\mathcal{A}$ *et un schéma* $s \in \mathcal{S}$ *tel que* $[(A_1, B_1), \ldots, (A_k, B_k)] \xrightarrow{s} (A, B)$ *?*

Notez que cette définition reste agnostique quant à la fa**c**con dont la relation de préférence est représentée dans les entrées.

Nous allons maintenant présenter dans la section suivante les différents schémas d'argument que nous pouvons trouver dans $\mathcal{S}$ et considérés pour le raisonnement sur les préférences.

3 Schémas de raisonnement sur les préférences

Cette section est consacrée à la construction de règles de dérivation adéquates pour raisonner sur les préférences. Nous formalisons ces règles comme des opérateurs liant une liste de prémisses à une conclusion, où prémisses et conclusions sont des comparaisons par paires. En faisant l'hypothèse d'un modèle additif, les schémas "*cancellation-based*" [2] ou une version simplifiée de ceux-ci, le schéma "*décomposition*", fournissent des règles qui sont correctes, mais la vérification de leur validité nécessite un calcul arithmétique qui peut sembler sans rapport avec la requête. Nous proposons d'éviter cette étape cognitivement insatisfaisante en identifiant les propriétés clés du modèle additif - la transitivité et la cancellation - et en formalisant des règles de dérivation tirant parti de chacune d'entre elles, à partir de la base : les schémas *transitif* et *ceteris paribus*. Nous introduisons ensuite le schéma *transitif réduit*, qui permet de dériver directement toute conclusion qui peut être prouvée à l'aide des deux schémas précédents. Ensuite, nous spécifions des exigences supplémentaires - *indépendance des éléments non pertinents (III)* [2] et commutation - qui donnent lieu à de nouveaux schémas, pour finalement aboutir au schéma *couverture*, qui particularise à la fois les schémas *transitif réduit* et *décomposition*.

3.1 Propriétés des préférences

Nous nous intéressons à la relation de préférence $\succsim$ qui pourrait exister entre les alternatives $A, B \in 2^{[m]}$, avec $A \succsim B$ signifiant que A est considéré comme au moins aussi bon que B.

2. Independance from Irrelevant Items

Nous rappelons quelques caractéristiques utiles que peuvent posséder les relations de préférence.

Définition 5 (Propriétés des préférences) *Soit* $\succsim \subset 2^{[m]} \times 2^{[m]}$ *une relation binaire entre alternatives. On dit :*

- $\succsim$ *est transitive lorsque, pour toute alternatives* $A, B, C \in 2^{[m]}$, *si* $A \succsim B$ *et* $B \succsim C$ *alors* $A \succsim C$ *;*
- $\succsim$ *satisfait cancellation (de premier ordre) si la préférence entre les alternatives ne dépend pas des éléments communs, c'est-à-dire* $\forall A, B \in 2^{[m]}\ A \succsim B \iff (A \setminus B) \succsim (B \setminus A)$ *;*
- $\succsim$ *est additif quand il existe un* m*-tuple de nombres réels* $\langle \omega_i \rangle_{i \in [m]} \in \mathbb{R}^{[m]}$ *tel que* $A \succsim B \iff \sum_{i \in A} \omega_i \geq \sum_{i \in B} \omega_i$.
- $\succsim$ *est un ordre linéaire additif lorsqu'il est additif et qu'il n'y a pas d'indifférence, c'est-à-dire que si* $A \neq B$ *donc soit* $A \not\succsim B$ *ou* $B \not\succsim A$ *[10].*

De toute évidence, une préférence additive satisfait à la fois aux propriétés de transitivité et de cancellation.

3.2 Le schéma transitif

Comme nous nous effor**c**cons d'expliquer les recommandations dérivant d'un modèle de somme pondérée, nous pouvons mécaniser les propriétés de transitivité et d'annulation sous la forme de règles de dérivation. Par exemple, nous définissons le *schéma transitif binaire* $(2 - tr)$, permettant de chaîner les affirmations de préférence comme suit :

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

En suivant l'approche que nous décrivons dans la section 2, étant donné un *explanandum* sous la forme d'une comparaison par paires (A, B) appartenant à $\succsim$, un *explanans* est une preuve consistant en des applications récursives d'une règle de dérivation – par exemple, *2-tr* – permettant de dériver la conclusion (A, B) à partir de prémisses acceptables (voir Section 2). Néanmoins, les preuves sont des objets récursifs qui peuvent être lourds à calculer ou à présenter au destinataire de l'explication, et nous proposons de pallier ce problème en introduisant des dispositifs de raisonnement plus puissants. En effet, considérons le cas d'un raisonnement purement transitif : enchaîner des lemmes transitifs revient à considérer des chaînes de prémisses transitives. Par exemple, si nous savons que $A \succsim B$, $B \succsim C$, $C \succsim D$ et $D \succsim E$, nous pouvons déduire que $A \succsim E$, en utilisant n'importe lequel des arbres de preuve décrits dans la Figure 1, et nous désignons $(A, B), (B, C), (C, D), (D, E) \vdash_{2-tr} (A, E)$. Nous pensons que cette abondance de preuves syntaxiques n'est pas pertinente pour la question du calcul des explications, et nous proposons donc de considérer le *schéma transitif* (tr) suivant.

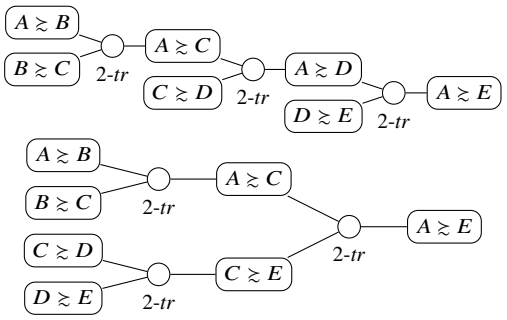


FIGURE 1 – Deux structures de preuve enchaînant des schémas transitifs binaires permettant de dériver la même conclusion à partir des mêmes axiomes.

Définition 6 (Schéma transitif (tr)) *La prémisse* $[(A_1, B_1), \ldots (A_k, B_k)]$ *et la conclusion* (A, B) *satisfont le* schéma transitif *quand, pour tout* $2 \leq j \leq k$, $A_j = B_{j-1}$, $A_1 = A$ *et* $B_k = B$.

Formellement, $\vdash_{2-tr} = \xrightarrow{tr}$ – *ce qui peut être prouvé en utilisant le schéma* $2 - tr$ *est exactement ce qui peut être dérivé en une seule application du schéma* tr. Nous avons échangé la nature récursive de la preuve utilisant un schéma binaire pour une dérivation *one-shot* utilisant un schéma opérant sur une liste de prémisses de longueur non limitée.

Exemple 1 *La prémisse* $[(\mathbf{acg}, \mathbf{bef}), (\mathbf{bef}, \mathbf{bfg})]$ *satisfait syntaxiquement le schéma transitif pour la conclusion* $(\mathbf{acg}, \mathbf{bfg})$ *:*

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

Qui peut être exprimé comme suit : "Dès lors que $\mathbf{acg} \succ \mathbf{bef}$ *et* $\mathbf{bef} \succ \mathbf{bfg}$, $\mathbf{acg}$ *devrait être préféré à* $\mathbf{bfg}$*". Notons toutefois que la première comparaison est complexe puisqu'il fait intervenir six critères différents.*

Notez que les comparaisons par paires composant la prémisse d'un schéma transitif sont ordonnées, donc la séquence d'alternatives $A \equiv A_0 \succsim B_0 \equiv A_1 \succsim \cdots \succsim B_{k-1} \equiv A_k \succsim B_k \equiv B$ est non croissante par rapport à la préférence.

Observation 1 *Si la préférence* $\succsim$ *est transitive, alors le schéma transitif est correct w.r.t.* $\succsim$ *: si toutes les comparaisons par pairess des prémisses sont dans* $\succsim$*, alors la conclusion est également une comparaison dans* $\succsim$.

Exemple 2 *Dans l'exemple 1, la prémisse appartient à* $\succsim$*, car pour* $(\mathbf{acg}, \mathbf{bef})$ *nous avons :* $242 = w_a + w_c + w_g > w_b + w_e + w_f = 219$*, et pour* $(\mathbf{bef}, \mathbf{bfg})$ *:* $219 = w_b + w_e + w_f > w_b + w_f + w_g = 204$*. Cependant, si nous utilisons la prémisse* $[(\mathbf{acg}, \mathbf{abc}), (\mathbf{abc}, \mathbf{bfg})]$*, cette dernière satisfait le schéma transitif mais n'appartient pas à* $\succsim$*, puisque pour* $(\mathbf{acg}, \mathbf{abc})$*, nous avons :* $242 = w_a + w_c + w_g < w_a + w_b + w_c = 331$ *(voir la fonction de score w dans la Section 2).*

3.3 Le schéma *ceteris paribus*

La propriété de cancellation permet de raisonner *ceteris paribus* – tout le reste étant égal – indépendamment du contexte, et représente une grande opportunité en termes d'explication, puisque la préférence peut être déduite de comparaisons par paires où les éléments communs ne sont pas mentionnés. Elle motive la définition suivante du schéma argumentatif *ceteris paribus*.

Définition 7 (Le schéma Ceteris paribus (cp)) *La prémisse* $[(A_1, B_1), \ldots (A_k, B_k)]$ *et la conclusion* (A, B) *satisfont le* schéma ceteris paribus *quand* $k = 1$, $A_1 \backslash B_1 = A \backslash B$ *et* $B_1 \setminus A_1 = B \setminus A$*. Dans ce cas, les comparaisons* (A_1, B_1) *et* (A, B) *sont dites* congruents.

De toute évidence, la congruence est une relation d'équivalence entre les comparaisons par paires. Dans la classe de congruence d'une comparaison donnée $(A, B) \in 2^{[m]} \times 2^{[m]}$, la comparaison $(A \setminus B, B \setminus A)$, où les alternatives sont disjointes par paires et obtenues à partir de (A, B) en soustrayant les éléments communs $A \cap B$ respectivement de A et B, présente un intérêt particulier. Lorsque (A, B) et (A_1, B_1) sont congruents, la préférence de A sur B se traduit par la préférence de $A \cap B$ sur $B \cap A$ en considérant que "tout le reste" – dans ce cas, les éléments de $A_c apB$ – "est égal" (ceteris paribus), puis à la préférence de A_1 sur B_1 en considérant $A_1 \cap B_1$ non pertinent.

Exemple 3 *En comparant* $\mathbf{acg}$ *à* $\mathbf{bfg}$*, il pourrait être justifié de considérer que, puisque* $\mathbf{g}$ *apparaît dans les deux alternatives, ce critère peut être omis : "la première option est meilleure que la seconde car, toutes choses égales par ailleurs,* $\mathbf{ac}$ *est préférable à* $\mathbf{bf}$*".*

Formellement, nous écrivons $[(\mathbf{ac}, \mathbf{bf})] \xrightarrow{cp} (\mathbf{acg}, \mathbf{bfg})$

Observation 2 *Si la préférence* $\succsim$ *satisfait la cancellation, alors le schéma* ceteris paribus *est correct en ce qui concerne* $\succsim$.

Nous serons souvent confrontés au problème inverse : étant donné un état initial, existe-t-il un état final tel que la comparaison par paires de l'état initial à l'état final est congruent à une comparaison par paires donnée ?

Lemme 1 (Quatrième problème de congruence) *Étant donnée une alternative* $A \in 2^{[m]}$ *et une comparaison par paires* $(A', B') \in 2^{[m]} \times 2^{[m]}$*, si* $A \supseteq (A' \setminus B')$ *et* $A \cap (B' \setminus A') = \emptyset$ *il existe exactement une alternative* $B \in 2^{[m]}$ *tel que* (A, B) *et* (A', B') *sont congruents, avec* $B = A \setminus (A' \setminus B') \cup (B' \setminus A')$*, sinon il n'en existe pas.*

Exemple 4 *Considérons la paire* $(\mathbf{ade}, \mathbf{bce})$*. L'ensemble des comparaisons par paires congruentes à celle-ci est :* $\{(\mathbf{ad}, \mathbf{bc}), (\mathbf{adf}, \mathbf{bcf}), (\mathbf{adg}, \mathbf{bcg}), (\mathbf{adef}, \mathbf{bcef}), (\mathbf{adeg}, \mathbf{bceg}), (\mathbf{adfg}, \mathbf{bcfg}), (\mathbf{adefg}, \mathbf{bcefg})\}$*. Les alternatives*

de départ {**ad**, **ade**, **adf**, **adg**, **adef**, **adeg**, **adfg**, **adefg**} *sont exactement les sur-ensembles de* **ad** *qui ne contiennent pas* **bc**, *et pour chacun d'entre eux, il n'existe qu'un seul état correspondant. Cela a du sens lorsque la comparaison pas paires* (**ade**, **bce**) *est comprise comme 'donner* **ade**, *prendre* **bce**', *dont* **e** *peut être omis. Ensuite,* **ad** *ne peut être effectivement pris que dans un état qui le contient déjà, et* **bc** *ne peut être effectivement ajouté que dans un état qui ne le contient pas encore.*

3.4 Le schéma transitif réduit

Lorsque la préférence satisfait à la fois aux propriétés de transitivité et de cancellation, il est correct d'utiliser à la fois les schémas tr et cp pour dériver de nouvelles comparaisons par paires.

La Figure 2 illustre une telle preuve. Pour des raisons pratiques, nous souhaitons rationaliser et mécaniser ce modèle de raisonnement, en formalisant un schéma - appelé *transitif réduit (rt)* - liant directement la prémisse $[(\mathbf{a},\mathbf{b}),(\mathbf{c},\mathbf{f})]$ à la conclusion $(\mathbf{aceg},\mathbf{befg})$ et en laissant les étapes intermédiaires non spécifiées [3].

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

FIGURE 2 – Une preuve de $[(\mathbf{a},\mathbf{b}),(\mathbf{c},\mathbf{f})] \vdash_{cp,tr} (\mathbf{aceg},\mathbf{befg})$

Définition 8 (Le schéma transitif réduit (rt)) $[(A_1,B_1),\ldots,(A_k,B_k)] \xrightarrow{rt} (A,B)$ *lorsqu'il existe* $(A'_1,B'_1),\ldots,(A'_k,B'_k)$ *and* (A',B') *tel que :*

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

Exemple 5 *La preuve représentée par la Figure 2 peut être lue comme suit :*

"On doit préférer **acg** *à* **bcg**, *puisque toutes choses étant égales par ailleurs,* **a** *est préféré à* **b**. *Alors* **bcg** *doit être préféré à* **bfg**, *puisque, toutes choses égales par ailleurs,* **c** *est préféré à* **f**. *Donc* **acg** *devrait être préféré à* **bfg**."

Nous notons :

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

3. Cela revient à permettre au schéma tr d'opérer sur l'ensemble quotient des classes congruentes de comparaisons par paires, au lieu d'être représentatif de ces classes.

Nous pouvons noter que la définition 8 est inefficace, car l'espace de recherche des séquences de comparaisons de longueur k est fini, mais intractablement grand. Le manque de spécification peut être surmonté en reconstruisant les étapes intermédiaires manquantes, en résolvant itérativement k quatrièmes problèmes congruents (voir Lemme 1), comme résumé par Algorithme 1 qui s'exécute en $O(km^2)$.

Le schéma rt remplit son rôle en permettant la dérivation en une seule étape de preuves combinant les dérivations tr et cp.

Proposition 1 *Si les prémisses* $\mathcal{P} = \bigcup_{j=1}^{n}(A_j,B_j)$ *permettent de prouver que la conclusion* (A,B) *via les schémas* cp *et* tr, *alors il existe une liste* $[P'_1,\ldots,P'_k]$ *de comparaisons par paires* $\mathcal{P}$ *tel que* $[P'_1,\ldots,P'_k] \xrightarrow{rt} (A,B)$.

Démonstration 1 *Par construction, le schémas* rt *subsume les schémas* tr *et* cp *(donc* $\vdash_{tr,cp} \subset \vdash_{rt}$*), et particularise une preuve combinant des dérivations de* tr *et* cp *(donc* $\vdash_{tr,cp} \supset \vdash_{rt}$*). Moreover, chaining* rt *derivations is useless, because if* $L_1 \xrightarrow{rt} (A,B)$ *et* $L_2 \xrightarrow{rt} (B,C)$, *alors* $L_1 \& L_2 \xrightarrow{rt} (A,C)$, *où* & *est la concaténation de liste. Du coup,* $\vdash_{rt} = \xrightarrow{rt}$.

Algorithm 1 checking an instance of the rt scheme

Require: a list of comparative statements $[(A_1,B_1),\ldots,(A_k,B_k)]$, a comparative statement $(A,\ B)$
Ensure: **true**, if the premise and conclusion satisfy the rt scheme ;
false, else
$A'_1 \leftarrow A \setminus B$
for $i = 1$ **to** k **do**
  if $A'_i \not\supseteq A_i$ **or** $A'_i \cap B_i \neq \emptyset$ **then**
    return False
  else
    $B'_i \leftarrow A'_i \setminus A_i \cup B_i$
    $A'_{i+1} \leftarrow B'_i$
  end if
end for
return $(B \setminus A = B'_k)$

3.5 Le schéma III - transitif réduit

Lors d'un raisonnement sur la préférence d'une alternative A par rapport à une autre alternative B, la mention d'un élément *neutre* j qui n'est ni dans A ni dans B, ou à la fois dans A et B, pourrait être considérée comme préjudiciable à l'explication. En effet, en raison de la propriété de cancellation, il pourrait être jugé inutile ; il pourrait être considéré comme hors sujet et vu comme un dispositif purement rhétorique ; il révèle des informations qui restent latentes lorsqu'on suit la stratégie *divulgation du modèle*. Nous appelons cette propriété *Indépendance des éléments non pertinents (Independence of Irrelevant Items –III)*, et nous l'appliquons de manière normative en considérant une

restriction du schéma transitif réduit aux instances qui la satisfont.

Définition 9 *Le schéma III- transitive réduit (III-rt)] Une prémisee* $[(A_1, B_1), \ldots, (A_k, B_k)]$ *et une conclusion* (A, B) *satisfont le* schéma III-transitive réduit *lorsqu'elle satisfont le schéma transitive réduit et* $\left(\bigcup_{i=1}^{k} A_i \cup \bigcup_{i=1}^{k} B_i\right) \subseteq (A \cup B) \setminus (A \cap B)$.

Exemple 6 *Le schéma dans l'Exemple 5 est un schéma III-transitive réduit. Cependant, si pour la même conclusion, nous utilisons la prémisse :* $[(\mathbf{acg}, \mathbf{bef}), (\mathbf{e}, \mathbf{g})]$*, le schéma ne l'est plus. En effet, les comparaisons par paires font intervenir le critère* $\mathbf{e}$ *qui n'apparaît pas dans la conclusion et peut sembler non pertinent.*

3.6 Le schéma décomposition

Introduit dans [2] et mettant en oeuvre des propriétés de cancellation d'ordre supérieur, le schéma décomposition vise à tirer parti de la propriété additive supposée de la relation de préférence [4]. Lorsque la préférence est additive, les affirmation de préférence se traduisent par des comparaisons linéaires, qui peuvent être additionnées. Ensuite, les scores des éléments apparaissant de part et d'autre s'annulent, permettant parfois de dériver de nouvelles comparaisons.

Définition 10 (Le schéma décomposition (*dec*)) *Une prémisse* $[(A_1, B_1), \ldots, (A_k, B_k)]$ *et une conclusion* (A, B) *satisfont le* schéma décomposition *lorsque chaque comparaison par paires* (A_i, B_i) *est disjointe et, pour touts les items* $j \in A \setminus B$*, il y a autant d'occurrences de* j *in the sets* $A_1, \ldots, A_k$ *qu'il y a dans les ensembles* $B_1, \ldots, B_k$ *plus un ; pour tous les items* $j \in B \setminus A$*, il y a autant d'occurrences de* j *dans les ensembles* $B_1, \ldots, B_k$ *qu'il y a dans les ensembles* $A_1, \ldots, A_k$ *plus un ; et pour tout élément* j *qui n'est ni dans A ni dans B, ou à la fois dans A et B, il y a autant d'occurrences de* j *dans les ensembles* $A_1, \ldots, A_k$ *que dans les ensembles* $B_1, \ldots, B_k$*, soit* $\forall\, j \in [m]$

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

Proposition 2 *Si la préférence* $\succsim$ *est additive, alors le schéma décomposition est correct par rapport à* $\succsim$.

Exemple 7 *Considérons le schéma décomposition suivant :*

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

En supposant que la préférence $\succsim$ *est additive, et que* $\mathbf{bc} \succsim \mathbf{de}$ *et* $\mathbf{efg} \succsim \mathbf{ac}$*. De la première comparaison, nous déduisons que* $\omega_{\mathbf{b}} + \omega_{\mathbf{c}} \geq \omega_{\mathbf{d}} + \omega_{\mathbf{e}}$*; de la seconde que* $\omega_{\mathbf{e}} + \omega_{\mathbf{f}} + \omega_{\mathbf{g}} \geq \omega_{\mathbf{a}} + \omega_{\mathbf{c}}$*. Par addition, nous obtenons* $\omega_{\mathbf{e}} + \omega_{\mathbf{f}} + \omega_{\mathbf{g}} + \omega_{\mathbf{b}} + \omega_{\mathbf{c}} \geq \omega_{\mathbf{d}} + \omega_{\mathbf{e}} + \omega_{\mathbf{a}} + \omega_{\mathbf{c}}$.

Ensuite, comme l'illustre la Figure 3 en annulant $\omega_{\mathbf{e}}$ *et* $\omega_{\mathbf{c}}$ *des deux côtés (il s'agit en fait d'une instance de* cacellation du second ordre*, car elle est effectuée sur deux comparaisons par paires), nous obtenons* $\omega_{\mathbf{f}} + \omega_{\mathbf{g}} + \omega_{\mathbf{b}} \geq \omega_{\mathbf{d}} + \omega_{\mathbf{a}}$*, alors* $\mathbf{bfg} \succ_{\omega} \mathbf{ad}$.

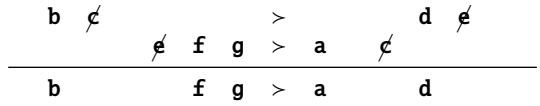


FIGURE 3 – Schéma décomposition : représentation graphique

Le schéma décomposition généralise strictement les schémas introduits précédemment.

Proposition 3 *Si la prémisse* $[(A_1, B_1), \ldots, (A_k, B_k)]$ *et la conclusion* (A, B) *satisfont le schéma transitif ou le schéma transitif réduit, alors elles satisfont le schéma décomposition.*

Nous notons que le schéma décomposition est commutatif par construction. Néanmoins, il ne semble pas très satisfaisant en tant que dispositif d'explication, car les propriétés de cancellation d'ordre supérieur qu'il met en oeuvre sont complexes et de faible intérêt normatif - en fait, même si elles sont syntaxiquement transductives, leur justification dérive de la forme additive que nous nous effor**c**cons de contourner.

En général, les instances du schéma décomposition ne satisfont pas le schéma transitif réduit. La séquence de comparaisons par paires de la prémisse ne peut pas être interprétée comme des justifications *ceteris paribus* de comparaisons par paires entre alternatives, car à un moment donné, ils nécessitent soit d'ajouter un élément à une alternative où il est déjà présent, soit de retirer un élément d'une alternative où il est absent.

Exemple 8 *La prémisse* $[(\mathbf{bc}, \mathbf{de}), (\mathbf{efg}, \mathbf{ac})]$ *et la conclusion* $(\mathbf{bfg}, \mathbf{ad})$ *satisfont le schéma décomposition, mais ne peuvent être interprétées comme une séquence de comparaisons par paires parce que par exemple l'alternative initiale* $\mathbf{bfg}$ *de la conclusion ne contient pas* $\mathbf{c}$.

Cette situation peut être évitée lorsque les alternatives mentionnées dans la prémisse sont toutes disjointes par pairess.

4. Ce schéma de décomposition est moins général que le schéma dit *syntactic cancellative* décrit dans [2], car il ne permet pas la répétition de la conclusion. Il a été démontré que cela réduit l'expressivité

Proposition 4 *Soient* $[(A_1, B_1), \ldots, (A_k, B_k)]$ *une prémisse et une conclusion* (A, B) *satisfaisant le schéma décomposition. Chaque permutation* σ *des indices* $[k]$ *pemet à la prémisse* $[(A_{\sigma(1)}, B_{\sigma(1)}), \ldots, (A_{\sigma(k)}, B_{\sigma(k)})]$ *et la conclusion* (A, B) *de satisfaire le schéma transitif réduit si et seulement si les alternatives* $A_1, \ldots, A_k, B_1, \ldots, B_k$ *sont disjointes par paires.*

Démonstration 2 *Supposons sans perte de généralité que* $A \cap B = \emptyset$*, et soit* j *apparaît dans au moins dans deux alternatives* $A_1, \ldots, A_k, B_1, \ldots, B_k$*.*

- *Supposons que* $j \notin A$*. S'il apparaît dans* A_i*, alors laissez* $\sigma(1) := i$*, sinon il apparaît dans* B_i *et* $B_{i'}$ *et donc* $\sigma(k-1) := i$ *et* $\sigma(k) := i'$*. Cela met en échec soit le premier ou le dernier quatrième problème congruent de Alg. 1.*
- *Supposons que* $j \notin B$*. S'il apparaît dans* B_i*, alors laissez* $\sigma(k) := i$*, sinon il apparaît dans* A_i *et* $A_{i'}$ *et donc* $\sigma(1) := i$ *et* $\sigma(2) := i'$*. Cela met en échec soir la vérification finale, soit le deuxième quatrième problème congruent de Alg. 1.*

Réciproquement, lorsque toutes les alternatives $A_1, \ldots, A_k, B_1, \ldots, B_k$ *sont disjointes par pairess, chaque élément de* $A \setminus B$ *apparaît exactement une fois dans le* B_i *et jamais dans le* A_i*, chaque élément de* $B \setminus A$ *apparaît exactement une fois dans le* A_i *et jamais dans le* B_i*, et les éléments des deux ensembles ou aucun n'apparaît jamais (ce qui rend incidemment la transaction III). Par conséquent, les éléments de l'ensemble* $\bigcup_{i=1}^{k} A_i$ *peuvent être retirés dans n'importe quel ordre de* A *et ceux de l'ensemble* $\bigcup_{i=1}^{k} B_i$ *peuvent être ajoutés dans n'importe quel ordre, de fa***c***con à s'accumuler dans* B*.*

Nous définissons en conséquence le *schéma couverture*.

3.7 Le schéma couverture

Dans ce schéma, une liste de comparaisons $[(A_1, B_1), \ldots, (A_k, B_k)]$ soutient une conclusion (A, B) si, et seulement si, les *pros* $A_1, \ldots, A_k$ partitionnent $A \setminus B$ et les *cons* $B_1, \ldots, B_k$ partitionnent $B \setminus A$.

Définition 11 (Le schéma couverture (cov)**)** *Une instance du* schéma couverture *est une instance du schéma décomposition où toutes les alternatives* $A_1, \ldots, A_k, B_1, \ldots, B_k$ *sont disjointes par paires.*

Le schéma couverture est commutatif et indépendant des éléments non pertinents par construction. Il particularise les schémas transitifs réduits – tout en contournant l'ordonnancement fastidieux des comparaisons par paires – et, comme corollaire, il est correct dans des conditions beaucoup plus souples que le schéma décomposition.

Proposition 5 *Si la préférence* $\succsim$ *est transitive et satisfait la cancellation, alors le schéma couverture est correct en ce qui concerne* $\succsim$*.*

Le schéma couverture décrit exactement l'algèbre morale introduite par Benjamin Franklin (voir [22]) pour déduire les préférences.

Exemple 9 *Considérons la conclusion :* (**bfg**, **cde**)*. La prémisse* [(**fg**, **e**), (**b**, **cd**)] *constitue un schéma couverture*

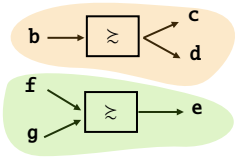
$$[(\mathbf{fg}, \mathbf{e}), (\mathbf{b}, \mathbf{cd})] \xrightarrow{cov} (\mathbf{bfg}, \mathbf{cde})$$

Le schéma couverture particularise à la fois le schéma transitif réduit et le schéma décomposition. En tant que tel, il nous donne le meilleur des deux mondes, en un sens. D'une part, il formalise une preuve, articulant des dérivations transitives et *ceteris paribus*, qui peut être présentée au destinataire de l'explication sous forme de diagramme, comme dans la Figure 4a, ou de manière narrative comme dans la Figure 4c (pour des comparaisons d'hôtels par exemple). D'un autre côté, les prémisses peuvent être comprises comme regroupant des contre avec des pour plus forts, de manière à "couvrir" les contre, et peuvent être présentées visuellement au destinataire de l'explication, comme dans la Figure 4b.

(a) Schéma couverture : diagramme de preuve de Ex. 9

$$\left.\begin{array}{l} \mathtt{fg} \succ \mathtt{e} \xrightarrow{cp} \mathtt{bfg} \succ \mathtt{be} \\ \mathtt{b} \succ \mathtt{cd} \xrightarrow{cp} \mathtt{be} \succ \mathtt{cde} \end{array}\right\} \xrightarrow{tr} \mathtt{bfg} \succ \mathtt{cde}$$

(b) Schéma couverture : une représentation visuelle de Ex.9



(c) Schéma couverture : une représentation narrative de Ex.9

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

FIGURE 4 – Trois représentations du schéma couverture

Notez également qu'un schéma couverture unique de longueur k peut être interprété comme $k!$ schémas transitifs, puisque la validité de ses prémisses ne dépend pas de leur ordre. Ainsi, par exemple pour notre Ex. 9, deux schémas transitifs correspondraient à ce schéma couverture : [(**bfg**, **cdfg**), (**cdfg**, **cde**)] or [(**bfg**, **be**), (**be**, **cde**)].

Résumé. Le Tableau 2 et la Figure 5 résument les schémas d'arguments introduits dans cette section, leurs propriétés, les exigences de la relation de préférence pour leur exactitude, et leurs relations.

Schéma	Propriétés	Exigences pour correction
décomposition	commutative	additivité
transitif réduit		transitivité + cancellation
III-transitif red.	III	transitivité + cancellation
covering	commutative, III	transitivité + cancellation
transitif.		transitivité
ceteris paribus		cancellation

TABLE 2 – Propriétés structurelles des schémas de raisonnement.

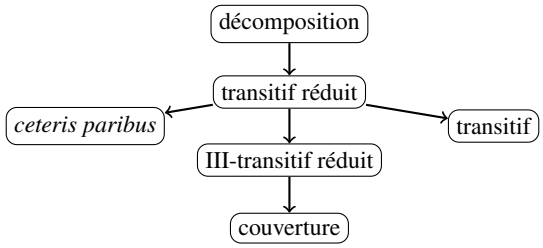


FIGURE 5 – Relations entre les schémas d'arguments. Une flèche allant de $scheme_1$ à $scheme_2$ indique que toutes les instances satisfaisant $scheme_2$ satisfont également $scheme_1$, mais pas l'inverse.

4 Expliquer avec des schémas

La Section 3 portait sur la conception d'un outil déductif permettant de dériver des préférences complexes à partir de préférences plus simples, d'une manière *correcte* par rapport au modèle de préférence latent. Étant donné un *explanans*–une comparaison par paires qui doit être expliquée–il nous permet de présenter le *problème d'explication* comme un problème d'*abduction* consistant à trouver des prémisses qui satisfont une certaine exigence de minimalité étant donné une conclusion et un ensemble de règles. Dans cette section, nous étudions l'expressivité relative et la complexité de calcul de l'explication avec les schémas *transitif réduite* (rt)–see Def. 8–et *couverture* (cov)–see Def. 11– ainsi que l'influence du choix des énoncés atomiquement simples.

À partir de maintenant, nous appellerons $\mathcal{E}(s, \mathcal{A}_{\succsim},)$ l'ensemble des pairess (A, B) qui peuvent être dérivées en utilisant le schéma s à partir de comparaison respectant les contraintes syntaxiques de $\mathcal{A}$, la contrainte sémantique d'être cohérent avec $\succsim$, et impliquant au plus k prémisses. Pour une paires donnée (A, B), le problème de l'existence de l'explication demande si $(A, B) \in \mathcal{E}(s, \mathcal{A}_{\succsim}, k)$. Par convention, cette paires (A, B) est considérée comme non évidente.

4.1 Résoudre les problèmes d'explication à l'aide de schémas

La figure 5 indique les dépendances logiques existant entre les prémisses et les conclusions satisfaisant les différents schémas. Cela a une conséquence évidente sur les relations d'explicabilité : plus général implique plus explicatif. De plus, les conclusions du schéma transitif peuvent toutes être obtenues via le schéma transitif réduit, et celles du *ceteris paribus* via les schémas décomposition, transitif réduit ou couverture.

Lorsque la préférence est additive, tous nos schémas de raisonnement sont corrects. Cependant, nous ne pouvons pas nous attendre à ce qu'ils soient *complets*, même sans aucune restriction syntaxique sur $\mathcal{A}$: lorsque $A \succ B$ sont adjacents dans la relation de préférence $\succsim$, c'est-à-dire lorsqu'il n'y a pas une autre alternative X s.t. $A \succ X \succ B$, (A, B) est appelée une *paires critique*, voir [10]. Mais cela signifie à son tour que la conclusion (A, B) ne peut pas être obtenue avec le schéma rt - encore moins par les schémas III-rt, cov et tr.

Ces pairess critiques ne sont donc pas explicables avec ces schémas (voir Ex.10).

Exemple 10 *La conclusion* $(\mathbf{bcd}, \mathbf{aefg})$ *constitue une paires critique. En effet,* $\omega_{\mathbf{b}} + \omega_{\mathbf{c}} + \omega_{\mathbf{d}} = 262$, $\omega_{\mathbf{a}} + \omega_{\mathbf{e}} + \omega_{\mathbf{f}} + \omega_{\mathbf{g}} = 258$, *et il n'est pas possible d'exhiber une autre alternative parmi les* $2^{[m]}$ *avec une note* $\in]258, 262[$.

En revanche, dès qu'une paires *n'est* pas critique, et à condition que $\mathcal{A}$ ne mette aucune contrainte syntaxique sur les comparaisons utilisées, il doit exister au moins une explication avec les schémas tr et rt. Cela signifie que la complexité de décider de l'existence d'une explication pour ces schémas est directement liée à celle de décider si une paires est critique. Nous montrons que ce problème est difficile.

Théorème 1 *Étant donné* $\omega \in \mathbb{N}_0^m$, *et* $A, B \in 2^{[m]}$ *tels que* $A \succsim B$ *où* $\succsim$ *est la relation de préférence additive induite de* ω. *Décider si* (A, B) *est une paires critique est Co-NP-complete.*

Démonstration 3 *Réduction de* SUBSET-SUM *[11].*

Dans SUBSET-SUM, *on nous donne un ensemble A de taille m, un poids positif* $w(a)$, *pour chaque a dansA, et un entier positif B. On se demande s'il existe un sous-ensemble* $A' \subseteq A$ *tel que la somme des poids soit exactement de K. Ce problème est connu pour être NP-complet. Nous construisons une instance du problème de la paires critique comme suit. On prend un ensemble C de* $m + 2$ *critères : pour chaque élément* $a_i \in A$, *on prend un critère* c_i, *de poids* $2s(a_i)$. *On ajoute deux autres critères :* a_{n+1} *de poids* $2K-1$, *et* a_{n+2}, *de poids* $2K+1$. *Nous demandons si la paires* (X, Y) *est critique, où* $X = \langle 0, \ldots, 1, 0\rangle$ *et* $Y = \langle 0, \ldots 1\rangle$. *Notons que X et Y ont des poids respectifs de* $2K - 1$ *et*

$2K+1$, *on cherche donc une alternative intermédiaire de poids exactement* $2K$, *un nombre pair. Nous affirmons que la réponse à cette question est non si le problème original de* SUBSET-SUM *est une instance de oui. Pour s'en convaincre, il suffit d'observer que tous les poids de notre instance de paires critique sont pairs, sauf ceux de* c_{n+1} *et* c_{n+2} *qui sont impairs. Comme* $s(c_{n+2}) > 2K$, *il ne peut certainement pas faire partie de la solution. De plus, la solution ne peut pas inclure* c_{n+1}, *car dans ce cas, il serait le seul poids impair, et la somme serait alors impaires. Nous nous retrouvons avec les critères* $c_1, \dots c_n$, *dont les poids sont précisément deux fois plus élevés que ceux du problème original de la somme des sous-ensembles.*

En corollaire, il est difficile de décider si une explication existe avec ces schémas.

Corollaire 1 *Étant donné* $\omega \in \mathbb{N}_0^m$ $A, B \in 2^{[m]}$ *tel que* $A \gtrsim B$, *où* $\gtrsim$ *est la relation de préférence additive induite par* ω, *et* $\mathcal{A}$ *l'ensemble des affirmations* $\Delta(m,m)$. *Décider si* $(A,B) \in \mathcal{E}(s, \mathcal{A}_\gtrsim, +\infty)$ *est NP-complete pour* $s \in \{rt, tr\}$.

Pour notre schéma de choix cov, nous avons le résultat suivant par une preuve indépendante.

Théorème 2 *Étant donné* $\omega \in \mathbb{N}_0^m$, $A, B \in 2^{[m]}$ *tel que* $A \gtrsim B$ *where* $\gtrsim$ *est une relation de préférence additive induite par* ω *et* $\mathcal{A}$ *l'ensemble des affirmations* $\Delta(m,m)$. *Décider si* $(A,B) \in \mathcal{E}(cov, \mathcal{A}_\gtrsim, +\infty)$ *est NP-complete.*

Démonstration 4 *(sketch) Membership is obvious, as the scheme itself is a polynomial certificate. Hardness results from reduction from* BIN-PACKING *[11].*

En ce qui concerne les autres schémas, alors que cp est facile, nous conjecturons que la complexité de dec et de III-rt est intractable.

4.2 Expliquer avec des affirmations atomiques

Nous abordons maintenant les explications qui imposent des restrictions syntaxiques aux ensembles d'éléments atomiques utilisés, $\Delta(1,1), \Delta(1,m),$ and $\Delta(m,1)$ (voir Sect.2).

Théorème 3 *Lorsque la relation* $\gtrsim$ *est additive,* $\mathcal{E}(cov, \mathcal{A}_\gtrsim, \infty)$ *est transitive lorsque* $\mathcal{A}_\gtrsim \in \{\Delta(1,1), \Delta(1,m), \Delta(m,1)\}$.

Démonstration 5 *Supposons* $[(A_1,B_1),\dots,(A_k,B_k)] \xrightarrow{cov} (A,B)$ *et* $[(A'_1,B'_1),\dots,(A'_k,B'_k)] \xrightarrow{cov} (B,C)$. *Nous montrons que la conclusion* (A,C) *est obtenu en appliquant le schéma couverture à une certaine prémisse. Il est facile de vérifier que* $[(A_1,B_1),\dots,(A_k,B_k),(A'_1,B'_1),\dots,(A'_k,B'_k)] \xrightarrow{dec} (A,C)$, *et nous avons seulement besoin de prouver que les ensembles* $A_1,\dots, A_k, A'_1, \dots, A'_{k'}, B_1,\dots, B_k, B'_1, \dots, B'_{k'}$, *sont disjoints par paires. Nous savos déjà que* $\langle A_i, B_{i'}\rangle$ *et que* $\langle A'_i, B'_{i'}\rangle$ *sont disjoint par paires. De plus,* A_i *sont inclus dans* $A \setminus B$ *tandis que* $A'_{i'}$ *sont dans* $B \setminus C$, *pour qu'ils ne se croisent pas (pareil pour* B_i *and* $B'_{i'}$*). Les seules intersections qu'il reste à considérer sont* $A_i \cap B'_{i'}$ *and* $B_i \cap A'_{i'}$. *Supposons sans perte de généralité* $A_i \cap B'_{i'} \neq \emptyset$. *En raison des contraintes syntaxiques* $\mathcal{A}$, *nous considérons cette intersection comme un singleton* $\{j\}$. *Nous supprimons les pairess* (A_i, B_i) *et* $(A'_{i'}, B'_{i'})$ *de la premisse, et les remplaccons par les pairess* $(A_i \cup A'_{i'} \setminus \{j\}, B_i \cup B'_{i'} \setminus \{j\})$. *Cette comparaison appartient à* $\mathcal{A}$, *et aussi à* $\gtrsim$ *car elle est additive (par addition des inégalités caractérisant* $A_i \gtrsim B_i$ *et* $A'_{i'} \gtrsim B'_{i'}$ *et annulation des termes* ω_j *apparaissant de part et d'autre). En itérant cette opération, on obtient un schéma couverture supportant* (A,C) *de taille non supérieure à* $k + k'$.

En corollaire, lorsque l'on restreint les atomes à la mention d'un seul pro contre un nombre quelconque de con (resp. un nombre quelconque de pro contre un con), le schéma transitif réduit n'est pas plus expressif que le schéma couverture. Ce n'est pas le cas lorsque nous permettons de mélanger ces atomes (comme illustré par Ex. 9).

Proposition 6 *Pour tout nombre entier positif* k, *lorsque* $\mathcal{A} \in \{\Delta(1,1), \Delta(1,m), \Delta(m,1)\}$ *et* $\gtrsim$ *est additive,* $\mathcal{E}(rt, \mathcal{A}_\gtrsim, k) = \mathcal{E}(cov, \mathcal{A}_\gtrsim, k)$.

On peut se demander si ces affirmations atomiques rendent le problème computationnelement plus simple à traiter. Bien que l'on sache que la réponse est positive pour $\mathcal{A} = \Delta(1,1)$ [1], nous montrons qu'à partir de $k \geq 2$ le problème est difficile.

Théorème 4 *Étant donné* $\omega \in \mathbb{N}_0^m$ $A, B \in 2^{[m]}$ *tel que* $A \gtrsim B$ *où* $\gtrsim$ *est une relation de préférence additive induite par* ω, *et* $\mathcal{A} = \Delta(1,k)$. *Lorsque* $k \geq 2$, *décider si* $(A,B) \in \mathcal{E}(cov, \mathcal{A}_\gtrsim, +\infty)$ *est NP-complete.*

Démonstration 6 *(sketch) Réduction de* 3D-MATCHING *[11].*

Lorsqu'on utilise le schéma couverture, la longueur des explications est limitée par le nombre m d'éléments. Par conséquent, il existe une gamme de valeurs de m pour lesquelles trouver une explication pourrait s'avérer trop difficile pour un humain, mais peut être facilement réalisé par une machine, soit avec un solveur ou même par force brute.

5 Complétude empirique du schéma couverture

Les résultats de la Section 4 établissent que les ensembles de comparaisons atomiques $\Delta(1,m)$ ou $\Delta(m,1)$ utilisant le schéma cov nous libèrent de la tâche de séquencer les

explications. Bien sûr, cela a un prix, car certaines pairess qui peuvent être expliquées autrement peuvent ne pas l'être avec ces comparaisons. Notre objectif est de donner un aperccu de la "complétude empirique" des comparaisons atomiques en utilisant le schéma cov.

Étant donné une relation de préférence additive spécifique $\mathcal{R}$, l'ensemble $\mathcal{A}_{\mathcal{R}}$ de comparaisons par paires qui peut être utilisé pour expliquer la paires (A, B) tel que $(A, B) \in \mathcal{R}$ et $(A, B) \notin \mathcal{A}_{\mathcal{R}}$ est le suivant :

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

lus précisément, nous considérons les relations de préférence sur les alternatives qui sont représentables par *un ordre linéaire additif sur l'algèbre des sous-ensembles d'un ensemble fini* (voir [10]). Nous supposons l'ordre suivant sur les alternatives singletons : $\mathbf{a} \succ \mathbf{b} \succ \mathbf{c} \succ \mathbf{d} \succ \mathbf{e} \succ \mathbf{f} \succ \ldots$. Alors, nous notons par $T^m_{\succ} = \{(A, B) \in 2^{[m]} \times 2^{[m]} : A \succ B \text{ and } A \cap B = \emptyset\}$, et nous avons $\mathcal{A}_{\succ} \subseteq T^m_{\succ}$. Le nombre d'ordres linéaires additifs sur $2^{[m]}$ croît très rapidement [16, 10] : il est de 14 pour $m = 4$, mais pour $m = 7$ nous avons déjà plus de 200 millions d'ordres.

Techniquement, nous avons été en mesure de générer touts les ordres linéaires additives de $m \in [[4; 6]]$. De plus, pour décider si $(A, B) \in \mathcal{E}(cov, \mathcal{A}_{\succ}, \infty)$, nous avons utilisé un programme linéaire mixte en nombres entiers. Enfin, pour évaluer les proportions de pairess $T^m_{\succ} \backslash \mathcal{A}_{\succ}$ qui sont explicables, on calcule pour chaque *ordre linéaire additif* $\succ$ donné m, la valeur suivante

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

m	Minimum	Médian	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 3 – $\mathfrak{M}_{m,\succ}$ for $m \in [[4; 6]] \quad \forall \succ$

Le Tableau 3 résume les valeurs minimale, médiane et maximale obtenues sur les ordres linéaires additifs respectivement de 14, 516 et 124187 (pour $m = 4, 5, 6$).

Nous remarquons que les valeurs minimale et médiane de $\mathfrak{M}_{m,\succ}$ augmentent avec m. En ce qui concerne les valeurs maximales, nous constatons qu'elles sont toutes égales à 100%, ce qui signifie que pour tous les $m \in [[4; 6]]$, il existe au moins un ordre linéaire additif $\succ$ pour lequel toutes les pairess sont explicables. En regardant plus globalement l'ensemble des valeurs du tableau 3, on peut dire qu'une majorité significative des pairess de $T^m_{\succ} \backslash \mathcal{A}_{\succ}$ sont explicables. Par exemple, pour $m = 6$, plus de 3 pairess sur 4 sont explicables quel que soit l'ordre linéaire additif considéré.

Bien sûr, l'explicabilité d'une paires arbitraire (A, B) dépend de ses caractéristiques par rapport au classement des critères qui les composent dans l'ordre sur les singletons. Par exemple, les pairess tels que (**ac**, **bd**) seront toujours explicables puisque $\mathbf{a} \succ \mathbf{b}$ et $\mathbf{c} \succ \mathbf{d}$. Cependant, il sera plus difficile de trancher pour des pairess comme (**ad**, **bc**) ou (**ae**, **bcd**) ou (**bde**, **acf**).

6 Travaux connexes

Récemment, [17] ont exploré les explications dans le contexte de la décision des classifieurs linéaires. Ils se concentrent sur les PI-explications (ou raisons suffisantes), c'est-à-dire les explications fournissant des raisons suffisantes pour expliquer une décision donnée, quelle que soit la valeur des autres critères [8, 7], une stratégie d'explication différente de la nôtre comme mentionné dans l'introduction. Notre vision des explications comme des preuves déductives cognitivement limitées rappelle les *systèmes de preuves limitées* proposés dans le contexte de la logique des descriptions [13, 9]. De même, une approche progressive similaire a été étudiée dans le contexte des problèmes de satisfaction de contraintes [3]. Enfin, les explications basées sur des axiomes ont été préconisées dans le domaine du choix social computationnel [5, 20]. En particulier, les travaux récents de [4] exploitent également les axiomes étudiés dans la théorie du vote pour produire des explications pour les décisions collectives, mais appliquées à un cadre différent (le vote), et en utilisant des techniques de preuve différentes (méthodes de tableau).

7 Conclusion

Nous proposons un cadre pour expliquer les comparaisons issues d'un modèle additif. Ce cadre s'accompagne de différents schémas pour raisonner sur les préférences, et nous nous concentrons sur un schéma spécifique : le schéma de couverture. De plus, à des fins cognitives, nous proposons de restreindre les explications aux ensembles d'éléments atomiques qui préfèrent un pour à un groupe de contre, ou un groupe de pour à un seul contre. Le moteur d'explication basé sur le recouvrement avec des ensembles restreints d'éléments atomiques n'est pas complet, mais des études empiriques montrent que des explications peuvent être calculées dans une grande proportion de cas.

De plus, le fait de fournir un schéma d'argumentation avec le résultat d'une comparaison par paires ouvre la possibilité de discuter ou de contester ce résultat. Ceci est rendu possible par ce qu'on appelle les questions critiques [21], un outil associé aux schémas d'argumentation représentant des attaques ou des critiques qui, si l'on n'y répond pas de manière adéquate, falsifient l'argument correspondant au schéma. Cela conduit naturellement à la perspective à long terme de la nature interactive du processus d'explication : ces schémas devraient être intégrés dans un processus dialectique, dans lequel l'utilisateur final devrait pouvoir

contester [19], tandis que d'autre part le système devrait acquérir des connaissances sur les préférences de l'utilisateur à travers un processus d'élicitation indirect [15]. L'imbrication harmonieuse de l'explication et de la recommandation appelle à la conception de systèmes d'initiative mixte [14] dans lesquels l'utilisateur peut être actif en défiant le système, et le système adaptatif dans ses réponses.

Références

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

[2] Belahcene, Khaled, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane: *Comparing options with argument schemes powered by cancellation*. In *Proc. 28th IJCAI*, pages 1537–1543, Macao, Macau SAR China, 2019.

[3] Bogaerts, Bart, Emilio Gamba, and Tias Guns: *A framework for step-wise explaining how to solve constraint satisfaction problems*. Artif. Intell., 300 :103–550, 2021.

[4] Boixel, Arthur, Ulle Endriss, and Ronald de Haan: *A Calculus for Computing Structured Justifications for Election Outcomes*. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI-2022)*, February 2022.

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

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

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

[8] Darwiche, Adnan and Auguste Hirth: *On The Reasons Behind Decisions*. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020.

[9] Engström, Fredrik and Claes Strannegård Abdul Rahim Nizamani: *Generating Comprehensible Explanations in Description Logic*. In *Informal Proceedings of the 27th International Workshop on Description Logics*, Vienna, 2014.

[10] Fishburn, Peter C., Aleksandar Pekec, and James A. Reeds: *subset comparisons for additive linear orders*. Mathematics of Operations Research, 27 :227–243, 2002.

[11] Garey, Michael R. and David S. Johnson: *Computers and Intractability, a Guide to the Theory of NP-completeness*. Freeman, 1979.

[12] Hammond, J, Ralph Keeney, and H Raiffa: *Even Swaps : A Rational Method for Making Trade-offs*. Harvard business review, 76 :137–8, 143, March 1998.

[13] Horridge, Matthew, Samantha Bail, Bijan Parsia, and Uli Sattler: *Toward Cognitive Support for OWL Justifications*. Know.-Based Syst., 53 :66–79, nov 2013, ISSN 0950-7051.

[14] Horvitz, Eric: *Uncertainty, Action, and Interaction : In Pursuit of Mixed-Initiative Computing*. Intelligent Systems, pages 17–20, 2000.

[15] Labreuche, Christophe, Nicolas Maudet, Wassila Ouerdane, and Simon Parsons: *A Dialogue Game for Recommendation with Adaptive Preference Models*. In *Proceedings of the 14th International Conference on Autonomous Agent and MultiAgent systems (AAMAS).*, pages 959–967, 2015.

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

[17] Marques-Silva, Joao, Thomas Gerspacher, Martin Cooper, Alexey Ignatiev, and Nina Narodytska: *Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay*. In Larochelle, H., M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (editors) : *Advances in Neural Information Processing Systems*, volume 33, pages 20590–20600. Curran Associates, Inc., 2020.

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

[19] Mulligan, Deirdre K., Daniel Kluttz, and Nitin Kohli: *Shaping Our Tools : Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*. In Werbach, Kevin (editor) : *After the Digital Tornado*. Cambridge University Press, 2020.

[20] Procaccia, Ariel D.: *Axioms Should Explain Solutions*. The Future of Economic Design, 2019.

[21] Walton, Douglas: *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996.

[22] Willcox, William B. (editor): *The Papers of Benjamin Franklin*, pages 299–300. New Haven and London, Yale University Press, 1975.