Différentiation des modalités du Bien : au-delà de l'optimalité de Pareto

Guillaume Gervois, Gauvain Bourgne, Marie-Jeanne Lesot

Différentiation des modalités du Bien : au-delà de l'optimalité de Pareto

Guillaume Gervois Gauvain Bourgne Marie-Jeanne Lesot

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
{prenom.nom}@lip6.fr

Résumé

L'éthique computationnelle étudie les restrictions et les préférences éthiques à intégrer aux algorithmes de prise de décision. Une approche pour faire face à une critique commune envers l'approche utilitariste de l'éthique computationnelle consiste à introduire des modalités différenciées du Bien, où les modalités sont définies comme les valeurs philosophiques qui correspondent aux différentes composantes du Bien. La différenciation permet alors qu'aucune modalité ne puisse en compenser une autre en définissant des classes distinctes de modalités. L'optimalité de Pareto modélise un cas extrême de différenciation, où chaque modalité constitue sa propre classe. Cet article propose une nouvelle approche, ordinale, pour traiter les modalités différenciées : la différenciation est modélisée par un ordre partiel strict sur les modalités, qui exprime quelles modalités prévalent sur les autres. L'article propose une axiomatisation de la supériorité pour prendre en compte ces comparaisons de modalités dans la détermination des actions éthiques : il discute de la manière d'induire une relation de préférence éthique entre les actions possibles, basée sur l'ordre partiel entre les modalités. En outre, il étudie les propriétés de cette relation induite, établissant qu'elle est asymétrique et transitive, prouvant ainsi qu'elle constitue une relation d'ordre.

Abstract

Computational ethics studies the ethical restrictions and preferences to be embedded into decision-making algorithms. One approach to deal with a common criticism towards the utilitarian approach to computational ethics consists in introducing differentiated modalities of the Good, where modalities are defined as philosophical values that correspond to different components of the Good. Differentiation then does not allow that any modality can compensate for any other one, distinct classes of modalities are defined. Pareto optimality models an extreme case of differentiation, where each modality constitutes its own class. This paper proposes a new, ordinal, approach to deal with differentiated modalities: differentiation is modelled by a strict partial order on the modalities, that expresses which modalities supersede others. The paper proposes an axiomatisation of superiority, to take into account these declared modality comparisons in the determination of ethical actions: it discusses how to derive an ethical preference relation between the possible actions, based on the partial order between the modalities. In addition, it studies the properties of this induced relation, establishing it is asymmetric and transitive, thus proving it constitutes a sound order relation.

1 Introduction

Les outils de prise de décision automatique sont de plus en plus répandus et utilisés. Face à cette popularité, on observe une demande croissante pour de nouveaux outils respectant les lois et les principes éthiques, c'est-à-dire vérifiant les contraintes de *conformité éthique*. Le domaine en pleine expansion de l'éthique computationnelle [1, 12] cherche à répondre à ces demandes. De nombreux principes éthiques proposés par des philosophes peuvent aider les informaticiens à aborder la question de la conformité éthique des algorithmes. L'utilitarisme, promu par Bentham et Mill à la fin du 18ème siècle, est l'une des théories morales les plus célèbres, mais aussi l'un des principes éthiques les plus implémentés [3, 10]. D'un point de vue computationnel, le principe utilitariste est séduisant car il est facilement représentable : il quantifie le Bien par des valeurs numériques, nommées utilités, qui peuvent ensuite être additionnées. Cependant, ce principe fait l'objet de débats philosophiques, notamment parce qu'il considère les différentes *modalités du Bien* comme étant toutes *indifférenciées*. Le terme *modalité* fait référence, ici et dans cet article, aux différentes valeurs philosophiques qui permettent de définir le Bien.

Prenons l'exemple d'une médecin dans un hôpital pour illustrer le fait que l'utilitarisme suppose que les modalités sont indifférenciées. Elle a le choix entre soigner un patient, ce que l'on note dans la suite *treat_patient* et qui aura pour effet de sauver une vie, et distribuer des chocolats à un

grand nombre de patients, noté *distribute_chocolat* et qui aura simplement pour effet de leur faire plaisir. Cet exemple confronte deux modalités : la vie humaine et le plaisir de manger du chocolat, notées respectivement *human_life* et *choco_pleasure*. Si l'on considère un nombre suffisamment important de patients, la somme des utilités attribuées au plaisir de manger du chocolat dépassera l'utilité attribuée au fait de sauver une vie, quelle que soit la valeur de cette dernière. L'utilitarisme conclut donc que la médecin doit distribuer du chocolat plutôt que de soigner le patient. Un tel cas montre que toute modalité peut être compensée par une autre : l'utilitarisme ne permet pas de modéliser la nature conflictuelle des modalités.

Les principales critiques de cette hypothèse d'indifférence font appel à une différenciation des modalités [8]. On peut par exemple considérer que le statut de médecin oblige à se préoccuper de la vie des patients plutôt que du plaisir de manger du chocolat, on peut aussi considérer que la vie humaine est plus importante que le plaisir de manger du chocolat. Cette dernière option introduit une notion de supériorité entre les modalités en accordant à certaines d'entre elles un statut particulier [5] : les modalités supérieures doivent être considérées en premier lorsqu'une décision doit être prise.

Suivant ces remarques, cet article propose une nouvelle approche, ordinale, pour traiter de la différentiation des modalités dans un système de conformité éthique : à notre connaissance, il propose une première tentative de relier cette préoccupation philosophique aux préférences ordinales. Plus précisément, il considère que la notion de supériorité est exprimée par un ordre partiel strict sur les modalités et il propose une *axiomatisation* de la supériorité, formalisant la prise en compte ces comparaisons de modalités afin d'en déduire des préférences ordinales entre les actions.

Le principe proposé peut être vu comme un principe de décision multicritère, où chaque modalité constitue un critère, allant au-delà du principe d'optimalité de Pareto : ce dernier, d'abord appliqué à des problèmes de prise de décision et ensuite à des problèmes éthiques [10], peut être considéré comme un cas extrême de différenciation des modalités. En effet, les modalités ne sont comparées qu'à elles-mêmes, et non les unes aux autres. Dans l'exemple médical précédent, aucune action n'est considérée comme dominant éthiquement l'autre : pour le principe de Pareto, les modalités sont incomparables entre elles. L'approche de supériorité que nous proposons et généralise le principe de Pareto en ajoutant la comparaison de supériorité des modalités.

L'article est structuré comme suit. La section 2 propose une formalisation du problème de conformité éthique afin de représenter les principes utilitariste et de Pareto, ainsi que la notion de comparaisons de modalité. La section 3 présente l'axiomatisation proposée de la supériorité qui prend en compte ces comparaisons pour déterminer une relation de préférence éthique entre les actions possibles. La section 4 étudie les propriétés de la relation induite proposée, établissant qu'elle constitue une relation d'ordre, prouvant qu'elle est asymétrique et transitive. La section 5 discute les hypothèses faites sur les relations de comparaison de modalité, au-delà du cas asymétrique et transitif. La section 6 conclut l'article et discute de certaines directions pour des travaux futurs.

2 Formalisation de la conformité éthique

Cette section décrit le formalisme considéré pour représenter un problème de conformité éthique, en présentant d'abord le cadre ordinal considéré et les notations utilisées tout au long de l'article. Elle introduit ensuite la représentation de la différenciation des modalités par un ordre partiel strict et montre enfin comment les principes utilitariste et de Pareto classiques sont exprimés dans ce cadre.

2.1 Formalisation ordinale de la conformité éthique

Un problème éthique consiste à sélectionner, parmi un ensemble $\mathcal{A}$ d'actions possibles (par exemple les options de soigner un patient ou distribuer du chocolat), l'ensemble $\mathcal{A}_p$ des *actions permissibles*, définies comme celles qu'il est éthiquement acceptable de réaliser selon un principe éthique donné. Dans l'article, les lettres a, a', o et o' seront utilisées pour représenter les éléments de $\mathcal{A}$.

Parmi les principes éthiques proposés par les philosophes et ceux qui ont été implémentés en éthique computationnelle, on retrouve l'utilitarisme de l'acte [13]. C'est une version courante de l'utilitarisme que l'on peut décomposer en trois étapes. Premièrement, les conséquences des actions sont éthiquement quantifiées par une *valeur d'utilité*. Dans la deuxième étape, ces valeurs d'utilité sont agrégées pour chaque action afin d'obtenir un nombre représentant l'utilité globale produite par l'action. Dans la dernière étape, les actions permissibles sont définies comme étant celles qui maximisent l'utilité.

Ces étapes peuvent être formalisées comme suit. Chaque action est représentée par un vecteur composé des valeurs d'utilité. Chaque valeur du vecteur correspond à une *modalité*, c'est-à-dire à l'une des valeurs philosophiques qui permettent de définir le Bien (par exemple la vie humaine ou le plaisir du chocolat). On note $\mathcal{M}$ l'ensemble fini des modalités et on considère que $\mathcal{A} \subset \mathbb{R}^{|\mathcal{M}|}$: plus la valeur du vecteur est élevée, plus l'action est intéressante du point de vue éthique selon cette modalité. Si l'action possède une valeur non nulle pour une modalité, on dit que l'action *porte* la modalité. Cette caractérisation des actions se situe dans le cadre usuel de la prise de décision multicritère [7], où les valeurs du vecteur peuvent être interprétées comme les

performances de l'action pour chacun des critères que sont les modalités.

Décrivons l'exemple présenté dans l'introduction avec ce formalisme.

Exemple 1. *Notons a l'action treat_patient et a' l'action distribute_chocolat. Considérons que sauver le patient a une valeur de* 10 *pour la modalité human_life, ainsi* $a_{human\_life} = 10$. *Ne procurant pas de plaisir aux patients, on a* $a_{choco\_pleasure} = 0$. *De même* $a'_{human\_life} = 0$. *Considérons que la distribution de chocolat procure* 1 *d'utilité et qu'il y a onze patients, soit* $a'_{choco\_pleasure} = 11$. *En utilisant la notation* $a = (a_{human\_life}, a_{choco\_pleasure})$, *cet exemple définit* $a = (10, 0)$ *et* $a' = (0, 11)$.

Cette quantification du Bien des conséquences est discutable : elle masque les relations causales en attribuant une seule valeur par modalité pour toutes les conséquences. On peut le voir directement avec l'action de distribuer du chocolat dans l'exemple 1. L'action telle qu'elle a été décrite cause un petit plaisir pour chacun des patients séparément. Elle possède donc un grand nombre de conséquences qui sont toutes valuées positivement pour la modalité *choco_pleasure*. Le formalisme proposé considère donc qu'une étape antérieure d'agrégation a déjà eu lieu, comme une somme pour l'utilitarisme de l'acte, afin de déterminer l'unique valeur de l'action *distribute_chocolat* pour la modalité *choco_pleasure*. En choisissant une autre fonction d'agrégation, il est possible de proposer d'autres solutions au problème du médecin. Ces solutions sont masquées par ce formalisme. Cependant, cette discussion dépasse le cadre souhaité dans cet article : la caractérisation choisie suffit à montrer l'intérêt d'une prise en compte différenciée des modalités.

Comme nous l'avons rappelé plus haut, l'utilitarisme de l'acte ordonne les actions en fonction de leurs utilités et définit comme permises celles qui ont les utilités les plus élevées. Pour formaliser cette vision ordinale, nous introduisons une relation de comparaison $\succsim_e$ sur $\mathcal{A} \times \mathcal{A}$ pour dénoter ces préférences éthiques. Ainsi $o \succsim_e o'$ signifie que l'action o est éthiquement préférée ou équivalente à l'action o'. La question est de savoir comment définir cette relation sur les actions, dont $\mathcal{A}_p$ est dérivé.

2.2 Différenciation ordinale des modalités

Comme discuté dans l'introduction, nous proposons de formaliser la différenciation des modalités comme un ordre partiel strict sur les modalités, que nous désignons par $\succ_m$, c'est-à-dire une relation asymétrique et transitive : $x \succ_m y$ signifie que la modalité x prévaut sur la modalité y. La modalité x est dite *dominante* et la modalité y *dominée*. L'ordre partiel strict peut être vu comme un ensemble de paires : $\succ_m \subset \mathcal{M}^2$. Chaque paire de modalités (x, y) est appelée une *comparaison*.

La difficulté de la définition de la supériorité consiste alors à prendre en compte ces comparaisons de modalités dans la détermination des actions admissibles : en ajoutant à la caractérisation des actions l'ordre partiel strict $\succ_m$ sur les modalités, il s'agit d'obtenir des informations sur la relation de comparaison $\succsim_e$, qui permettra ensuite d'obtenir l'ensemble $\mathcal{A}_p$.

Notre formalisation de la supériorité entre modalités a pour objectif de modéliser le fait qu'aucun plaisir issu de la consommation de chocolat, aussi grand qu'il soit, ne peut jamais égaler ou dépasser le fait de sauver une vie. Autrement dit de modéliser la préférence pour les actions qui portent une modalité dominante plutôt que n'importe laquelle des actions ne portant que des modalités dominées. De ce fait, la formalisation ne fournit que des relations de préférence stricte entre les actions, et pas de relations d'équivalence. Nous nous intéressons donc particulièrement à la partie asymétrique de $\succsim_e$ qui est notée $\succ_e$, où $o \succ_e o'$ signifie que o est strictement préférée à o'.

2.3 Formalisation des principes éthiques classiques

Dans cet article, la relation de préférence est définie à l'aide de propriétés de la forme suivante :

$$\forall o, o' \in \mathcal{A}, [\textit{conditions sur } o,\ o' \textit{et les modalités}] \Rightarrow o \succ_e o' \quad (1)$$

2.3.1 Utilitarisme de l'acte

Le principe de l'utilitarisme de l'acte rappelé dans la section précédente peut être exprimé comme suit :

$$\forall o, o' \in \mathcal{A},\ \left[\sum_{x \in \mathcal{M}} o_x > \sum_{x \in \mathcal{M}} o'_x \right] \Rightarrow o \succ_e o' \quad (2)$$

On obtient pour l'exemple 1 : $\sum_{x \in \mathcal{M}} a_x = 10$ et $\sum_{x \in \mathcal{M}} a'_x = 11$. Selon l'équation 2, l'utilitarisme de l'acte conclut $a' \succ_e a$.

2.3.2 Optimalité de Pareto

Le principe de Pareto classique utilisé dans le cadre de la décision multicritère peut être écrit dans sa version stricte comme suit :

$$\forall o, o' \in \mathcal{A},\ \left[\exists x \in \mathcal{M},\ (o_x > o'_x) \wedge (\forall y \in \mathcal{M} \backslash \{x\},\ o_y \geq o'_y)\right] \Rightarrow o \succ_e o' \quad (3)$$

Pour l'exemple 1, on observe $a_{human\_life} > a'_{human\_life}$ et $a'_{choco\_pleasure} > a_{choco\_pleasure}$. Les deux modalités ne favorisant pas la même action, l'optimalité de Pareto ne fournit aucune préférence.

2.3.3 Discussion

Les deux principes précédents assurent la transitivité et l'asymétrie de $\succ_e$. En terme de traitement des modalités, le principe utilitariste considère que les modalités sont équivalentes, puisque dans l'équation 2 la somme est une fonction d'agrégation commutative : on peut inverser les modalités sans modifier le résultat. Au contraire, le principe de Pareto considère que les modalités sont incomparables : dans l'équation 3, seules les quantifications d'une même modalité sont comparées entre les actions considérées.

La contribution de cet article, telle que décrite dans les sections suivantes, se concentre sur la définition d'une nouvelle condition plus expressive que ces deux cas extrêmes pour le traitement des comparaisons entre modalités. Elle combine les quantifications par modalités et l'ordre $\succ_m$ entre les modalités afin d'introduire la supériorité entre les modalités.

3 Proposition d'axiomatisation de la supériorité entre les modalités

Cette section décrit la définition proposée d'une relation de préférence éthique entre les actions possibles, basée sur l'ordre partiel entre les modalités, résultant en une axiomatisation de la supériorité, comme une généralisation du principe de Pareto. Elle formalise d'abord la définition du comportement de supériorité souhaité, puis décrit en trois étapes l'axiomatisation proposée, en fonction du nombre de modalités dominantes et dominées.

3.1 Définition de la supériorité

Afin de définir le comportement de supériorité souhaité, nous considérons d'abord le cas où l'ordre partiel strict sur la modalité contient une seule comparaison, notée $x \succ_m z$. La supériorité de la modalité x sur la modalité z est alors définie dans le formalisme par l'équivalence suivante :

$$x \succ_m z \Leftrightarrow \big[\forall o, o' \in \mathcal{A},\ [(o_x > o'_x \wedge \forall y \in \mathcal{M}\backslash\{x,z\},\ o_y \geq o'_y)] \Rightarrow o \succ_e o'\big] \quad (4)$$

Le point important de cette définition est que les quantifications de la modalité dominante x sont suffisantes pour déterminer la préférence entre deux actions, indépendamment des quantifications de la modalité dominée. Il n'y a donc pas de compensation possible entre une modalité dominante et une modalité dominée. Quant aux autres modalités y qui ne sont pas impliquées dans la comparaison, comme pour l'optimalité de Pareto, il est nécessaire qu'elles favorisent la même action que la modalité dominante.

3.2 One Over One : un dominant, un dominé

Dans le cas où l'ensemble de comparaison définit une seule modalité dominante et une seule modalité dominée, la définition de la relation $\succ_e$ induite découle directement de la définition de la supériorité de l'équation 4 :

$$\forall o, o' \in \mathcal{A}, \big[\exists x, z \in \mathcal{M},\ x \succ_m z \wedge (o_x > o'_x) \wedge (\forall y \in \mathcal{M}\backslash\{x,z\},\ o_y \geq o'_y)\big] \Rightarrow o \succ_e o' \quad (5)$$

Reprenons l'exemple 1 en y ajoutant la comparaison $human\_life \succ_m choco\_pleasure$. On sait que $a_{human\_life} > a'_{human\_life}$. Ne disposant que de deux modalités dans cet exemple, la condition sur les y est vérifiée également. Ainsi, l'équation déduit la préférence $treat\_patient \succ_e distribute\_chocolat$.

3.3 One Over Many : un dominant, plusieurs dominés

Dans un problème complexe, on peut être amené à considérer un ensemble de comparaisons. Cette section considère le cas où une seule modalité dominante prévaut sur un ensemble de modalités dominées. Dans ce cas, nous considérons la généralisation suivante de l'équation 5 : quel que soit le nombre de modalités dominées, elles ne peuvent pas contrer la préférence induite par la modalité dominante. Cette généralisation est une supposition forte qui donne à la propriété de supériorité proposée un caractère *absolu* : rien ne peut la contredire.

$$\forall o, o' \in \mathcal{A}, \big[\exists x \in \mathcal{M}, \exists Z \subset \mathcal{M}\backslash\{x\}, (\forall z \in Z,\ x \succ_m z) \wedge (o_x > o'_x) \wedge (\forall y \in \mathcal{M}\backslash\{x\} \cup Z,\ o_y \geq o'_y)\big] \Rightarrow o \succ_e o' \quad (6)$$

Cette propriété est équivalente à la reformulation suivante :

$$\forall o, o' \in \mathcal{A}, \big[\exists x \in \mathcal{M}, (o_x > o'_x) \wedge \forall y \in \mathcal{M}\backslash\{x\}, (x \succ_m y \vee o_y \geq o'_y)\big] \Rightarrow o \succ_e o' \quad (7)$$

Cette dernière formule souligne le fait qu'elle peut être considérée comme une généralisation de l'optimalité de Pareto. En effet, si aucune comparaison n'est considérée, alors $x \succ_m y$ est faux pour toutes les modalités et la formule est identique à l'équation 3.

3.4 Many Over Many : cas général

Dans le cas général, pour toute paire d'actions o et o', il faut distinguer trois sous-ensembles de modalités de $\mathcal{M}$:

— L'ensemble X des modalités favorisant une même action, qui doit être non vide afin d'obtenir une préférence stricte en favorisant une action o par rapport à une action o' :

$$X = \{x \in \mathcal{M} \mid o_x > o'_x\}$$

— L'ensemble des modalités dominées, qui représente les modalités dominées par au moins une modalité de l'ensemble X :

$$\{y \in \mathcal{M} \backslash X \mid \exists x \in X,\ x \succ_m y\}$$

— L'ensemble des modalités non dominantes et non dominées, qui doivent être en accord avec les modalités de l'ensemble X :

$$\{y \in \mathcal{M} \backslash X \mid o_y \geq o'_y\}$$

Par rapport au cas précédent, cette généralisation renforce le caractère *absolu* de la supériorité en précisant que la présence d'une seule modalité dominante suffit à considérer une modalité comme étant dominée. Une modalité dominée ne participe activement que si aucune préférence n'est exprimée pour toutes ses modalités dominantes. Dans ce cas, nous proposons la définition suivante :

$$\begin{aligned} &\forall o, o' \in \mathcal{A}, \\ &\quad \big[\exists X \subset \mathcal{M},\ X \neq \emptyset,\ (\forall x \in X,\ o_x > o'_x) \wedge \\ &\quad [\forall y \in \mathcal{M} \backslash X,\ (\exists x \in X,\ x \succ_m y) \vee o_y \geq o'_y]\big] \\ &\qquad\qquad \Rightarrow o \succ_e o' \end{aligned} \quad (8)$$

Cette propriété est équivalente à la reformulation suivante :

$$\begin{aligned} &\forall o, o' \in \mathcal{A},\ \big[\exists x \in \mathcal{M}, (o_x > o'_x)\ \wedge [\forall y \in \mathcal{M}, \\ &\quad (\exists x' \in \mathcal{M}, x' \succ_m y \wedge o_{x'} > o'_{x'}) \vee o_y \geq o'_y]\big] \\ &\qquad\qquad \Rightarrow o \succ_e o' \end{aligned} \quad (9)$$

Comme pour le cas précédent, il s'agit d'une généralisation de l'optimalité de Pareto. Si aucune comparaison n'est considérée, alors $x' \succ_m y$ est faux pour toutes les modalités et l'équation 9 est identique à l'équation de l'optimalité de Pareto.

3.5 Définition de la préférence minimale induite $\succ_e^m$

Parmi l'ensemble de toutes les préférences $\succ_e$ qui satisfont l'équation 9, la relation de préférence minimale est définie comme celle qui ne contient que les paires induites par l'équation. Ainsi pour définir cette relation, il suffit de remplacer l'implication de l'équation 9 par une équivalence.

Définition 1. *La* préférence éthique minimale*, notée $\succ_e^m$, est la relation de préférence induite uniquement par l'équation 9 :*

$$\begin{aligned} &\forall o, o' \in \mathcal{A}, \\ &\exists x \in \mathcal{M}, (o_x > o'_x) \wedge \\ &[\forall y \in \mathcal{M}, (\exists x' \in \mathcal{M}, x' \succ_m y \wedge o_{x'} > o'_{x'}) \vee o_y \geq o'_y] \\ &\Leftrightarrow o \succ_e^m o' \end{aligned}$$

En utilisant cette définition, un ordre $\succ_e$ satisfait l'axiomatisation de la supériorité que nous proposons dans l'équation 9 si et seulement si il est un sur-ensemble de cette relation minimale : $\succ_e^m \subseteq \succ_e$.

La section suivante étudie les propriétés de cette relation de préférence minimale, en établissant qu'elle est asymétrique et transitive, ce qui implique que c'est un ordre partiel strict.

4 Propriétés de la relation $\succ_e^m$ proposée

Cette section établit que la relation $\succ_e^m$ proposée satisfait la propriété requise de définir une relation d'ordre sur les actions :

Théorème 1. *$\succ_e^m$ est un ordre partiel strict.*

Les sections 4.1 et 4.2 prouvent respectivement qu'il est asymétrique et transitif. Les deux preuves utilisent le lemme suivant où $\oplus$ désigne le XOR binaire :

Lemme 1. *Pour tout ensemble non vide $X \subseteq \mathcal{M}$, en notant l'ensemble des modalités maximales* $\max_{\succ_m}(X) = \{x \in X \mid \forall x' \in X,\ \neg(x' \succ_m x)\}$*, on a :*

$$\forall x \in X, (x \in \max_{\succ_m}(X)) \ \oplus\ (\exists x' \in \max_{\succ_m}(X),\ x' \succ_m x)$$

Démonstration. Ce lemme est prouvé par récurrence sur $|X|$.

— Si $|X| = 1$, alors $X = \{x\} = \max_{\succ_m}(X)$.

— Si $|X| = n+1$, avec $n \in \mathbb{N}^*$. On a $X = X' \cup \{x\}$, avec $|X'| = n$. Dans ce cas, on distingue deux possibilités :

  — $x \in \max_{\succ_m}(X)$.

  — $x \notin \max_{\succ_m}(X)$, par définition de $\max_{\succ_m}(X)$, on a $\exists x' \in X,\ x' \succ_m x$. D'après l'asymétrie de $\succ_m$, on peut conclure que $x \neq x'$ d'où $x' \in X'$. Par hypothèse de récurrence sur X' on obtient soit $x' \in \max_{\succ_m}(X')$, et on pose $x'' = x'$, soit $\exists x'' \in \max_{\succ_m}(X'),\ x'' \succ_m x'$. Par transitivité et asymétrie, on a $x'' \succ_m x$ et $\neg(x \succ_m x'')$. Donc on obtient $x'' \in \max_{\succ_m}(X)$ et $x'' \succ_m x$.

□

4.1 Asymétrie de la relation $\succ_e^m$ proposée

Proposition 1. *$\succ_e^m$ est asymétrique :*
elle vérifie $\forall o, o' \in \mathcal{A},\ o \succ_e^m o' \Rightarrow \neg(o' \succ_e^m o)$.

Démonstration. On suppose que $o \succ_e^m o'$ et par absurde que $o' \succ_e^m o$. D'après la définition 1, on obtient :

— $\exists x_0 \in \mathcal{M},\ (o_{x_0} > o'_{x_0})$ (A)

— $\forall y \in \mathcal{M}$,
$o_y \geq o'_y \vee (\exists x' \in \mathcal{M},\ x' \succ_m y \wedge o_{x'} > o'_{x'})$ (B)

— $\exists x_1 \in \mathcal{M},\ (o'_{x_1} > o_{x_1})$ (C)

— $\forall y \in \mathcal{M}$,
$o'_y \geq o_y \vee (\exists x' \in \mathcal{M},\ x' \succ_m y \wedge o'_{x'} > o_{x'})$ (D)

Appelons S l'ensemble des modalités qui ont une préférence pour o plutôt que o' et I l'ensemble des modalités qui ont une préférence pour o' plutôt que o.

$S = \{x \in \mathcal{M} \mid o_x > o'_x\}$ et $I = \{x \in \mathcal{M} \mid o'_x > o_x\}$. D'après (A) et (B), on sait que ces ensembles sont non vides. S étant non vide et en utilisant le Lemme 1, on peut prendre un $z \in \max_{\succ_m}(S)$. Ainsi, $z \in S$ donc $o_z > o'_z$ et donc, avec (D), $\exists x_2 \in \mathcal{M},\ x_2 \succ_m z \wedge o'_{x_2} > o_{x_2}$. $o'_{x_2} > o_{x_2} \Rightarrow x_2 \in I$ et en utilisant le Lemme 1 sur I :

— si $x_2 \in \max_{\succ_m}(I)$, on note $x_3 = x_2$.

— sinon $\exists x_3 \in \max_{\succ_m}(I),\ x_3 \succ_m x_2$.

Dans les deux cas on obtient $x_3 \succ_m z$ par transitivité de $\succ_m$. $x_3 \in I$ donc $o'_{x_3} > o_{x_3}$ et avec (B), $\exists x_4 \in \mathcal{M}, x_4 \succ_m x_3 \wedge o_{x_4} > o'_{x_4}$. On déduit $o_{x_4} > o'_{x_4}$ donc $x_4 \in S$. Par transitivité $x_4 \succ_m z$, de plus par définition de $\max_{\succ_m}(S)$, $x_4 \in S$ et $x_4 \succ_m z$ donc $z \notin \max_{\succ_m} S$, ce qui est contradictoire. On conclut donc que $\neg(o \succ_e^m o' \wedge o' \succ_e^m o)$. □

4.2 Transitivité de la relation $\succ_e^m$ proposée

Proposition 2. *$\succ_e^m$ est transitive :*
elle vérifie $\forall o, o' \in \mathcal{A}$,

$$(o \succ_e^m o' \wedge o' \succ_e^m o'') \Rightarrow (o \succ_e^m o'').$$

Démonstration. Considérons o, o', o'' tel que $o \succ_e^m o'$ et $o' \succ_e^m o''$. En utilisant la définition 1, on obtient :

— $\exists x_0 \in \mathcal{M},\ (o_{x_0} > o'_{x_0})$ (E_1)

— $\forall y \in \mathcal{M},\ o_y < o'_y \Rightarrow (\exists x' \in \mathcal{M},\ x' \succ_m y \wedge o_{x'} > o'_{x'})$ (E_2)

— $\exists x_1 \in \mathcal{M},\ (o'_{x_1} > o''_{x_1})$ (F_1)

— $\forall y \in \mathcal{M},\ o'_y < o''_y \Rightarrow (\exists x' \in \mathcal{M},\ x' \succ_m y \wedge o'_{x'} > o''_{x'})$ (F_2)

On doit prouver $o \succ_e^m o''$, soit d'après la définition 1, $P_1 : \exists x \in \mathcal{M}, o_x > o''_x$, et pour tout $y \in \mathcal{M}$,
$P_2(y) : o_y < o''_y \Rightarrow \exists z \in \mathcal{M}.\ z \succ_e^m y \wedge o_z > o''_z$.

Preuve de P_1. Par E_1, on a x_0 tel que $o_{x_0} > o'_{x_0}$. Si $o'_{x_0} \geq o''_{x_0}$ alors $o_{x_0} > o''_{x_0}$ et P_1 est satisfait. Sinon, $o'_{x_0} < o''_{x_0}$. Selon le lemme 1 et F_2, $S_0 = \{x \in \mathcal{M} | x \succ_m x_0 \wedge o'_x > o''_x\}$ est non vide, ainsi on peut choisir x_2 dans $max_{\succ_m} S_0$. Si $o_{x_2} \geq o'_{x_2}$ alors $o_{x_2} > o''_{x_2}$ et P_1 est satisfait. Sinon, $o_{x_2} < o'_{x_2}$. Par E_2, on obtient une modalité x_3 telle que $x_3 \succ_m x_2$ et $o_{x_3} > o'_{x_3}$ Comme x_2 est maximale pour $\succ_m$ dans S_0, on a $x_3 \notin S_0$ et donc $o'_{x_3} \leq o''_{x_3}$. Si $o'_{x_3} < o''_{x_3}$, utiliser F_2 donnerait une modalité de S_0 supérieure à x_2, ce qui contredirait sa maximalité. Donc $o'_{x_3} = o''_{x_3}$. Avec $o_{x_3} > o'_{x_3}$, cela implique P_1.

Preuve de $\forall y, P_2(y)$. Considérons une modalité $y_0 \in \mathcal{M}$. Si $o_{y_0} \geq o''_{y_0}$, $P_2(y_0)$ est trivialement vérifié. Sinon, on a $o_{y_0} < o''_{y_0}$ (H_1). On a donc deux cas :

(A) Supposons $o_{y_0} < o'_{y_0}$ (H_2). Selon le lemme 1, E_2 et H_2, $S_2 = \{x \in \mathcal{M} | x \succ_m y_0 \wedge o_x > o'_x\}$ est non vide, ainsi on peut choisir x' dans $max_{\succ_m} S_2$.

(A.1) Supposons $o'_{x'} < o''_{x'}$ (H_3). Par F_2 et H_3, on obtient une modalité z telle que $z \succ_m x' \wedge o'_z > o''_z$. On a $z \succ_m x'$ et $x' \succ_m y_0$, donc, par transitivité de $\succ_m$, $z \succ_m y_0$. Etant donné que x' est maximal pour $\succ_m$, on doit avoir $z \notin S_2$ ce qui donne $o_z \geq o'_z$. Avoir $o_z > o'_z$ n'est pas possible car cela autoriserait à dériver depuis E_2 une modalité qui appartiendrait à S_1 tout en étant supérieure à x', contredisant encore la maximalité de x'. On peut conclure $o_z = o'_z$, et donc $o_z > o''_z$, ce qui prouve $P_2(y_0)$.

(A.2) Sinon, $o'_{x'} \geq o''_{x'}$. Etant donné un $x' \in S_1$, on obtient $o_{x'} > o''_{x'}$. On a ainsi (prenant x' pour z), $P_2(y_0)$.

(B) Dans l'autre cas, $o_{y_0} \geq o'_{y_0}$ (H_4). On considère ensuite les modalités qui sont supérieures à y_0.

(B.1) Supposons que $\exists y' \in \mathcal{M}, y' \succ_m y_0 \wedge o_{y'} < o'_{y'}$. Alors, en appliquant le raisonnement du cas A.1 à y', on obtient un $z \in \mathcal{M}$ tel que $z \succ_m y'$ et $o_z > o''_z$. Par transitivité de $\succ_m$, $z \succ_m y_0$, ce qui prouve $P_2(y_0)$.

(B.2) Sinon, on doit avoir : $\forall y' \in \mathcal{M}, y' \succ_m y_0 \Rightarrow o_{y'} \geq o'_{y'}$ (H_5). Avec H_1 et H_4, on a $o'_{y_0} < o''_{y_0}$. Appliquer F_2 donne une modalité z telle que $z \succ y_0$ et $o'_z > o''_z$. Etant donné H_5 on a $o_z \geq o'_z$ et donc $o_z > o''_z$, ce qui prouve $P_2(y_0)$.

Nous avons ainsi prouvé $P_2(y_0)$ dans tous les cas et pour tout y_0. □

Ceci conclut la preuve du théorème 1. $\succ_e^m$ est un ordre partiel strict.

5 Discussions sur les hypothèses faites sur $\succ_m$

Nous avons supposé que la relation $\succ_m$ est asymétrique et transitive, cela englobe de nombreuses situations, néanmoins nous discutons dans cette section deux cas alternatifs.

5.1 Cas d'une relation $\succ_m$ totale

L'ajout d'autres hypothèses peut donner des informations supplémentaires sur $\succ_e$. Par exemple, si nous supposons que la relation $\succ_m$ est également totale, alors l'axiomatisation devient un ordre lexicographique sur les modalités [6]. Il suffit alors pour départager les actions d'observer s'il existe une préférence pour la modalité au sommet de l'ordre, et ainsi de suite jusqu'à la fin de l'ordre $\succ_m$. Ainsi, pour toute paire d'actions non égales o et o', une préférence sera déduite de l'équation 9. Cette propriété est utile si l'on souhaite une action unique à réaliser. Cependant, le fait d'avoir une seule action permise peut être vu comme une propriété restrictive pour un système de conformité éthique.

5.2 Cas d'une relation $\succ_m$ non transitive

On peut aussi souhaiter que la relation $\succ_m$ ne soit pas transitive. Cependant, cette section montre que c'est une condition nécessaire à notre axiomatisation si l'on souhaite

définir des préférences rationnelles entre différentes actions. En effet, si on autorise une relation $\succ_m$ qui n'est pas transitive, il est alors possible de définir des boucles de supériorité entre modalités. Supposons par exemple que l'on dispose de quatre modalités w, x, y, z telles que $w \succ_m x$, $x \succ_m y$, $y \succ_m z$ et $z \succ_m w$. Dans ce cas, l'axiomatisation proposée dans l'équation 9 ne garantit plus l'asymétrie et la transitivité de la relation minimale induite $\succ_e^m$. Pour l'illustrer, considérons deux actions o et o' telles que $o_w > o'_w$, $o_x < o'_x$, $o_y > o'_y$ et $o_z < o'_z$. Appliquons maintenant l'équation 9 :

— Pour obtenir $o \succ_e^m o'$: on observe $o_w > o'_w$ donc $\exists x \in \mathcal{M}, (o_x > o'_x)$ est vérifiée. De plus on vérifie $\forall y \in \mathcal{M}, o_y \geq o'_y$ pour les modalités w et y. Quant aux modalités x et z, il existe w et y telles que $w \succ_m x \wedge o_w > o'_w$ et $y \succ_m z \wedge o_y > o'_y$. Ainsi x et z vérifient $\forall y \in \mathcal{M}, (\exists x' \in \mathcal{M}, x' \succ_m y \wedge o_{x'} > o'_{x'})$. Toutes les conditions sont vérifiées, on en déduit $o \succ_e o'$.

— Pour obtenir $o' \succ_e^m o$, on applique exactement le même raisonnement mais en partant de la modalité x au lieu de la modalité w.

La relation induite $\succ_e^m$ n'est donc pas asymétrique. On peut alors observer des cycles dans les préférences obtenues entre les différentes actions. Ces cycles dans les préférences posent différents problèmes [9], comme l'argument de la pompe monétaire, qui empêchent de les considérer comme rationnelles. Si l'on souhaite éviter les cycles tout en conservant une relation $\succ_m$ qui n'est pas transitive, il est nécessaire de modifier l'axiomatisation proposée. Néanmoins, de telles modifications sortent du cadre voulu pour cet article étant donné qu'elles nécessitent d'introduire des conditions ne provenant pas directement du concept de supériorité entre modalités.

6 Conclusion et perspectives

Cet article propose une axiomatisation du concept philosophique de supériorité entre les modalités du Bien. Pour ce faire, un formalisme de décision multicritère ordinal adapté à la prise de décision éthique a été défini, basé sur une approche utilitariste. En tant que généralisation du principe d'optimalité de Pareto, l'axiomatisation proposée permet de déduire les préférences à partir de la différenciation des modalités.

Le travail présenté dans cet article ouvre de multiples perspectives. Il constitue une première étape pour relier les préoccupations philosophiques aux préférences ordinales. Les travaux en cours visent à étudier les cadres formels existants qui offrent des propriétés similaires à celles que nous proposons, comme par exemple les hiérarchies de contraintes [2], les CP-nets et TCP-nets [4] ou encore des variantes des méthodes de surclassement avec seuil [11].

Une limite du travail actuel réside dans les simplifications faites sur les relations causales dans le formalisme utilisé, comme discuté dans la section 2. Afin de prendre en compte les questions éthiques qui interviennent sur ces relations causales, nous envisageons d'étendre le formalisme pour pouvoir prendre en compte chaque conséquence des actions séparément.

Comme discuté dans la section 5, l'ensemble minimal de préférences éthiques qui respectent un ordre de supériorité $\succ_m$ entre les modalités n'est pas nécessairement total. En effet, le principe de supériorité qui est axiomatisé n'a pas pour objectif de résoudre toutes les décisions éthiques. Cela soulève donc des questions sur la combinaison de plusieurs principes afin d'obtenir un unique ensemble d'actions permissibles. Ainsi, des travaux en cours ont pour objectif de formaliser une version plus générale du concept de principe éthique et des conditions que le mélange de plusieurs principes doit respecter.

Références

[1] Anderson, Michael et Susan Leigh Anderson: *Machine Ethics*. Cambridge University Press, 2011.

[2] Borning, Alan, Bjorn Freeman-Benson et Molly Wilson: *Constraint hierarchies*. LISP and symbolic computation, 5(3) :223–270, 1992.

[3] Bourgne, Gauvain, Camilo Sarmiento et Jean Gabriel Ganascia: *ACE modular framework for computational ethics : dealing with multiple actions, concurrency and omission*. Dans *1st Workshop on Computational Machine Ethics*, 2021.

[4] Brafman, Ronen I, Carmel Domshlak et Solomon Eyal Shimony: *On graphical modeling of preference and importance*. Journal of Artificial Intelligence Research, 25 :389–424, 2006.

[5] Chang, Ruth: *Incommensurability (and Incomparability)*. John Wiley & Sons, Ltd, 2013.

[6] Fishburn, Peter C.: *Axioms for Lexicographic Preferences*. The Review of Economic Studies, 42(3) :415–419, 1975.

[7] Gonzales, Christophe et Patrice Perny: *Multicriteria Decision Making*, page 519–548. Springer International Publishing, 2020.

[8] Griffin, James: *Are There Incommensurable Values ?* Philosophy & Public Affairs, 7(1) :39–59, 1977.

[9] Hansson, Sven Ove et Till Grüne-Yanoff: *Preferences*. Dans *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2022.

[10] Lindner, Felix, Martin Mose Bentzen et Bernhard Nebel: *The HERA approach to morally competent robots*. Dans *2017 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 6991–6997. IEEE, 2017.

[11] Rogers, Martin, Michael Bruen et Lucien Yves Maystre: *The Electre Methodology*, pages 45–85. Springer US, Boston, MA, 2000.

[12] Tolmeijer, Suzanne, Markus Kneer, Cristina Sarasua, Markus Christen et Abraham Bernstein: *Implementations in Machine Ethics : A Survey*. ACM Comput. Surv., 53(6), décembre 2021.

[13] Vallentyne, Peter: *Consequentialism*. Dans *Philosophy publications*. Wiley-Blackwell, 2006.