Dealing with differentiated modalities of the Good: beyond Pareto optimality

Guillaume Gervois, Gauvain Bourgne, Marie-Jeanne Lesot

Dealing with differentiated modalities of the Good: beyond Pareto optimality

Guillaume Gervois[0000−0002−7425−5490], Gauvain Bourgne[0000−0002−5104−9352], and Marie-Jeanne Lesot[0000−0002−3604−6647]

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

Abstract. 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.

Keywords: computational ethics · utilitarianism · ethical preferences

1 Introduction

Automatic decision making tools are becoming more and more complex. They call for tools allowing to verify that they respect laws and ethical principles, within the growing field of machine ethics [1, 8]. Many ethical principles have been proposed by philosophers that can help computer scientists to address this issue of ethical compliance of algorithms. Utilitarianism, promoted by Bentham and Mill at the end of the 18th century, is one of the most famous moral theories, but also one of the most implemented ethical principles [3, 7]. Indeed, from a computational point of view, the utilitarian principle is attractive because it is easily representable: it quantifies the Good with numerical values, named utilities, that are then summed up. However, this principle is the subject of philosophical debates, notably because it conceals the notion of *modalities of the Good*. The term *modality* refers, here and in this paper, to the different philosophical values which allow to define the Good.

Consider the example of a physician in a hospital to illustrate the fact that utilitarianism assumes that the modalities are equivalent. He or she has the

choice between treating a patient, which will result in saving a life, and distributing chocolates to a large number of patients, which will simply result in pleasing them. This example confronts two modalities: human life and the pleasure of eating chocolate. If a sufficiently large number of patients is considered, the sum of the utilities assigned to the pleasure of eating chocolate will exceed the utility assigned to saving a life, whatever the value of the latter. Thus utilitarianism will conclude that the physician should distribute chocolate rather than treating the patient. Such a case shows that any modality can be compensated by another one: utilitarianism does not take into account the conflicting nature of the modalities.

The main criticisms of this equivalence assumption appeal to a differentiation of modalities [6]. Some argue that the status of physician implies that he or she should care for the lives of patients rather than the pleasure of eating chocolate, others argue that human life is more important than the small pleasure of eating chocolate. This last criticism introduces a notion of superiority between the modalities by granting some of them a special status [4]: superior modalities must be considered first when a decision has to be made.

Following these lines, this paper proposes a new, ordinal approach, to deal with differentiated modalities within an ethical compliance system: it proposes a first attempt, to the best of our knowledge, to connect this philosophical concern with ordinal preferences. More precisely, it considers that the notion of superiority is expressed by a strict partial order on the modalities and it proposes an axiomatisation of superiority, to take into account these modality comparisons to compare the actions and make an ethical choice.

The proposed principle can be seen as going beyond the Pareto optimality principle: the latter, first applied for decision making problems and then also used in ethical ones [7] can be considered as an extreme case of modality differentiation. Indeed, modalities are compared only to themselves, and not to each other. In the previous example of the physician, no action is then considered as ethically dominating the other: for the Pareto principle, the modalities are incomparable to each other. The proposed superiority approach generalises this property, supplementing it with modality superiority comparison.

This paper is structured as follows. Section 2 proposes a formalisation of ethical compliance problem in order to represent the utilitarian and the Pareto principles, as well as the notion of modality comparisons. Section 3 presents the proposed axiomatisation of superiority that takes into these comparisons to determine an ethical preference relation between the possible actions. Section 4 studies the properties of the proposed induced relation, establishing it constitutes an order relation, proving it is asymmetric and transitive. Section 5 discusses the assumptions made on the modality comparison relations, beyond the asymmetric and transitive case. Section 6 concludes the paper and discusses some directions for future works.

2 A Formalisation of Ethical Compliance

This section describes the considered formalism for representing an ethical compliance problem, first presenting the considered ordinal framework and the notations used throughout the paper. It then introduces the representation of modality differentiation through a strict partial order and finally shows how the classical utilitarian and Pareto principles are expressed in this framwework.

2.1 An Ordinal Formalisation of Ethical Compliance

An ethical problem consists in selecting, among a set $\mathcal{A}$ of possible actions (e.g. treating a patient or distributing chocolate), the set $\mathcal{A}_p$ of the *permissible actions*, defined as the ones that are ethically acceptable to perform according to a given ethical principle [9].

Among the possible ethical principles that have been proposed by philosophers and implemented in the machine ethics domain, act utilitarianism [9], a common version of utilitarianism, can be decomposed into three steps. First, the consequences of actions are ethically quantified by a *utility value*. In the second step, these utility values are aggregated for each action in order to obtain a number representing the global utility produced by the action. In the final step, the permissible actions are defined as the ones that maximise utility.

These steps can be formalised as follows. Each action is represented by a vector, composed of the aggregation of the utility values of its consequences. Each vector component corresponds to a *modality*, that corresponds one of the philosophical values which allow to define the Good (e.g. human life or chocolate pleasure). We note $\mathcal{M}$ the considered finite set of modalities and consider the case where $\mathcal{A} \subset \mathbb{R}^{|\mathcal{M}|}$: the higher the vector component, the more ethically interesting the action according to that modality. With this characterisation of actions, the modalities can be interpreted as criteria in multi-criteria decision making. This quantification of the Good of consequences is debatable, as it hides causal relations by assigning a single value per modality for all consequences. However, this discussion is beyond the scope of this paper and this characterisation is sufficient to show the interest of a differentiated consideration of modalities.

As recalled above, act utilitarianism orders actions based on their utilities and defines the ones with highest utilities as permissible. To formalise this ordinal view, we introduce a comparison relation $\succsim_e$ to denote these ethical preferences. The question is how to define this relation on the actions, from which $\mathcal{A}_p$ is derived.

2.2 Ordinal Differentiation of the Modalities

As discussed in the introduction, we propose to formalise modality differentiation as a strict partial order on the modalities, which we denote $\succ_m$, i.e. an asymmetric and transitive relation: $x \succ_m y$ means that modality x supersedes modality y. Modality x is said *dominant* and modality y *dominated*. The strict

partial order can be seen as a set of pairs: $\succ_m \subset \mathcal{M}^2$. Each pair of modalities (x, y) is called a *comparison*.

The issue of defining superiority then consists in taking into account these declared modality comparisons in the determination of the permissible actions: by adding to the characterisation of the actions a strict partial order on the modalities, the aim is to obtain information on the pre-order relation $\succsim_e$, which will then allow to obtain the set $\mathcal{A}_p$. We can already notice that superiority will not provide equivalence information between the actions, but strict preference information. We are therefore particularly interested in the asymmetric part of the preorder $\succsim_e$ which is noted $\succ_e$.

2.3 Formalisation of Classical Ethical Principles

Formally, the global aim of implementing ethical principles is to define the preference order $\succsim_e$ among possible actions, usually inferred from is asymmetric part $\succ_e$. In this paper, it will be defined using properties of the following form:

$$\forall o, o' \in \mathcal{A}, [\textit{some constraints on } o,\ o' \textit{ and modalities}] \Rightarrow o \succ_e o'$$

Act Utilitarianism The act utilitarian principle recalled in the previous section can be expressed as follows:

$$\forall o, o' \in \mathcal{A},\ \left[\sum_{x \in \mathcal{M}} o_x > \sum_{x \in \mathcal{M}} o'_x \right] \Rightarrow o \succ_e o' \tag{1}$$

Pareto Optimality The classical Pareto principle used in multi-decision framework can be expressed with the property of increasing monotonicity, whose strict version can be written as

$$\forall o, o' \in \mathcal{A},\ \left[\exists x \in \mathcal{M},\ (o_x > o'_x) \wedge (\forall y \in \mathcal{M} \backslash \{x\},\ o_y \geq o'_y)\right] \Rightarrow o \succ_e o' \tag{2}$$

Discussion The two previous principles ensure the transitivity and asymmetry of $\succ_e$. The utilitarian principle considers that the modalities are equivalent, since in Eq.1 the sum is a commutative aggregation function. On the contrary, the Pareto principle considers the modalities are incomparable: in Eq. 2, only quantifications of the same modality are compared for the considered actions.

The contribution of this paper, as described in the next sections, focuses on defining a new constraint to introduce superiority among modalities by combining the quantifications of consequences by modalities and the $\succ_m$ order among modalities.

3 Proposed Axiomatisation of Superiority Among Modalities

This section describes the proposed definition of an ethical preference relation between the possible actions, based on the partial order between the modalities,

resulting in an axiomatisation of superiority, as a generalisation of the Pareto principle. It first formalises the definition of the desired superiority behaviour and then describes in three steps the proposed definition, depending on the number of dominant and dominated modalities.

3.1 Definition of Superiority

In order to define the desired superiority behaviour, we first consider the case where the strict partial order on the modality contains a single comparison, denoted $x \succ_m y$. Superiority is then defined as

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

The important point of this definition is that quantifications of the dominant modality are sufficient to determine the preference between two actions regardless of the quantifications of the dominated modality. There is therefore no compensation possible between a dominant and a dominated modality.

Considering the medical example discussed in the introduction, with the comparison *human_life* $\succ_m$ *chocolate_pleasure*, , whatever the number of patients receiving chocolat, this model leads to *treat_patient* $\succ_e$ *distribute_chocolat*.

Like increasing monotonicity, this superiority is conditioned by the value of the other modalities, those that are neither dominant nor dominated.

3.2 One Over One: Single Dominant, Single Dominated

In the case where the comparison set defines a single dominant modality and a single dominated modality, the definition of the induced $\succ_e$ follows from the superiority definition recalled in the previous section:

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

3.3 One Over Many: Single Dominant, Many Dominated

In a complex problem, one may have to consider a set of comparisons. This section considers the case where a single dominant modality supersedes a set of dominated modalities. In this case, we consider the following generalisation of Eq. 4: whatever the number of dominated modalities, they cannot counter the preference induced by the dominant modality.

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

This property is equivalent and therefore reformulated as follows:

$$\forall o, o' \in \mathcal{A}, \exists x \in \mathcal{M}, (o_x > o'_x) \wedge (\forall y \in \mathcal{M}\backslash\{x\}, x \succ_m y \vee o_y \geq o'_y) \Rightarrow o \succ_e o' \quad (6)$$

The latter formula underlines that it can be seen as a generalisation of the increasing monotony. Indeed, if no comparison is considered, then $x \succ_m y$ is false for all modalities and the formula is identical to that of the increasing monotonicity.

3.4 Many Over Many: General Case

In the general case, for any pair of actions o and o', three subsets of modalities of $\mathcal{M}$ must be distinguished

- The set of dominant modalities X, which must be non-empty to obtain a strict preference and thus favour an action o over an action o':
$$X = \{x \in \mathcal{M} \mid o_x > o'_x\}$$
- The set of dominated modalities, which represents the modalities dominated by at least one dominant modality:
$$\{y \in \mathcal{M} \backslash X \mid \exists x \in X,\ x \succ_m y\}$$
- The set of non-dominant and non-dominated modalities, which must be in agreement with the dominant modalities:
$$\{y \in \mathcal{M} \backslash X \mid o_y \geq o'_y\}$$

In this case, we propose the following definition:

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

This property is equivalent and therefore reformulated as follows:

$$\forall o, o' \in \mathcal{A},\ \exists x \in \mathcal{M}, (o_x > o'_x) \wedge \\ [\forall y \in \mathcal{M}, (\exists x' \in \mathcal{M}, x' \succ_m y \wedge o_{x'} > o'_{x'}) \vee o_y \geq o'_y] \Rightarrow o \succ_e o' \quad (8)$$

If no comparison is considered, then $x' \succ_m y$ is false for all modalities and the formula is identical to that of the increasing monotony. The general case of the axiomatisation we proposed, given in Eq. 8, is therefore a generalisation of the Pareto principle.

3.5 Definition of the Minimal Preference Relation $\succ_e^m$

Among the set of all $\succ_e$ preferences that satisfy Eq. 8, the minimal preference relation is defined as the one that contains only the pairs induced by the equation:

Definition 1. *The minimal ethical preference, noted $\succ_e^m$, is the preference relation induced only by Eq. 8:*

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

Using this definition, an order $\succ_e$ satisfies the axiomatisation of superiority we propose in Eq. 8 if and only it is a superset of this minimal relation: $\succ_e^m \subseteq \succ_e$. The next section studies the properties of this minimal preference relation, establishing that it is both asymetric and transitive, which means that it is a strict partial order.

4 Properties of the Proposed $\succ_e^m$ Relation

This section establishes that the proposed $\succ_e^m$ relation satisfies the required property of defining an order relation on the actions:

Theorem 1. *$\succ_e^m$ is a strict partial order.*

Sections 4.1 and 4.2 respectively prove that it is asymmetric and transitive. Both proofs use the following lemma:

Lemma 1. *For any non-empty set $X \subseteq \mathcal{M}$ and the set of maximal modalities $\max_{\succ_m}(X) = \{x \in X \mid \forall x' \in X,\ x' \neg\succ_m x\}$, it holds that*

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

Proof. We are going to prove this lemma by recurrence on $|X|$.

- If $|X| = 1$, then $X = \{x\}$ and by definition of $\max_{\succ_m}(X)$, $x \in \max_{\succ_m}(X)$.
- If $|X| = n + 1$, with $n \in \mathbb{N}^*$. We have $X = X' \cup \{x\}$, with $|X'| = n$. In that case, either:
  - $x \in \max_{\succ_m}(X)$.
  - $x \notin \max_{\succ_m}(X)$, by definition of $\max_{\succ_m}(X)$, we get $\exists x' \in X,\ x' \succ_m x$. With $\succ_m$ asymmetry, we can conclude that $x \neq x'$ and $x' \in X'$. By recurrence hypothesis on X' we get either $x' \in \max_{\succ_m}(X')$, and we note $x'' = x'$, or $\exists x'' \in \max_{\succ_m}(X'),\ x'' \succ_m x'$. By transitivity and asymmetry, we get $x'' \succ_m x$ and $x \neg\succ_m x''$. Therefore we have $x'' \in \max_{\succ_m}(X)$ and $x'' \succ_m x$.

4.1 Asymmetry of the Proposed $\succ_e^m$ Relation

Proposition 1. *$\succ_e^m$ is asymmetric: it verifies $o \succ_e^m o' \Rightarrow \neg(o' \succ_e^m o)$.*

Proof. We suppose that $o \succ_e^m o'$, and by absurd, let's suppose that $o' \succ_e^m o$. By using Def. 1, we get:

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

Let's call S the set of modalities which have a preference for o over o' and I the set of modalities which have a preference for o' over o. $S = \{x \in \mathcal{M} \mid o_x > o'_x\}$ and $I = \{x \in \mathcal{M} \mid o'_x > o_x\}$. With (A) and (B), we know that these sets are not empty. S being non empty and by using Lemma 1, we can take $z \in \max_{\succ_m}(S)$. Thus, $z \in S \Rightarrow o_z > o'_z$ and $o_z > o'_z \wedge (D) \Rightarrow \exists x_2 \in \mathcal{M},\ x_2 \succ_m z \wedge o'_{x_2} > o_{x_2}$. $o'_{x_2} > o_{x_2} \Rightarrow x_2 \in I$ and by using Lemma 1 on I:

- $x_2 \in \max_{\succ_m} I$ and we say that $x_3 = x_2$.
- $\exists x_3 \in \max_{\succ_m} I,\ x_3 \succ_m x_2$, and by transitivity of $\succ_m$, we get $x_3 \succ_m z$.

$x_3 \in I \Rightarrow o_{x_3} > o'_{x_3}$ and $o'_{x_3} > o_{x_3} \wedge (B) \Rightarrow \exists x_4 \in \mathcal{M},\ x_4 \succ_m x_3 \wedge o_{x_4} > o'_{x_4}$. $o_{x_4} > o'_{x_4} \Rightarrow x_4 \in S$. By transitivity $x_4 \succ_m z$, however by definition of $\max_{\succ_m}(S)$, $x_4 \in S \wedge x_4 \succ_m z \Rightarrow z \notin \max_{\succ_m} S$. It is absurd, thus we conclude that $\neg(o \succ_e^m o' \wedge o' \succ_e^m o)$.

4.2 Transitivity of the Proposed $\succ_e^m$ Relation

Proposition 2. *$\succ_e^m$ is transitive: it verifies $(o \succ_e^m o' \wedge o' \succ_e^m o'') \Rightarrow (o \succ_e^m o'')$.*

Proof. Let's consider o,o',o'' such that $o \succ_e^m o'$ and $o' \succ_e^m o''$. By using Def. 1, we get:

- $\exists x_0 \in \mathcal{M},\ (o_{x_0} > o'_{x_0})\ (E_1)$
- $\forall y \in \mathcal{M},\ o_y < o'_y \Rightarrow (\exists x' \in \mathcal{M},\ x' \succ_m y \wedge o_{x'} > o'_{x'})\ (E_2)$
- $\exists x_1 \in \mathcal{M},\ (o'_{x_1} > o''_{x_1})\ (F_1)$
- $\forall y \in \mathcal{M},\ o'_y < o''_y \Rightarrow (\exists x' \in \mathcal{M},\ x' \succ_m y \wedge o'_{x'} > o''_{x'})\ (F_2)$

We have to prove $o \succ_e^m o''$, that is (given Def. 1), $P_1 : \exists x \in \mathcal{M}, o_x > o''_x$ and for all $y \in \mathcal{M}$, $P_2(y) : o_y < o''_y \Rightarrow \exists z \in \mathcal{M}. z \succ_e^m y \wedge o_z > o''_z$.

Proof of P_1. By E_1, we have x_0 such that $o_{x_0} > o'_{x_0}$. If $o'_{x_0} \geq o''_{x_0}$ then $o_{x_0} > o''_{x_0}$ and P_1 is satisfied. Otherwise, $o'_{x_0} < o''_{x_0}$. By Lemma 1, F_2, $S_0 = \{x \in \mathcal{M} | x \succ_m x_0 \wedge o'_x > o''_x\}$ is not empty, so we can choose x_2 in $max_{\succ_m} S_0$. If $o_{x_2} \geq o'_{x_2}$ then $o_{x_2} > o''_{x_2}$ and P_1 is satisfied. Otherwise, $o_{x_2} < o'_{x_2}$. From E_2, we get a modality x_3 such that $x_3 \succ_m x_2$ and $o_{x_3} > o'_{x_3}$ Since x_2 maximal for $\succ_m$ in S_0, we have $x_3 \notin S_0$ and thus $o'_{x_3} \leq o''_{x_3}$. But if $o'_{x_3} < o''_{x_3}$, applying F_2 would give a a modality of S_0 superior to x_2 which would contradict its maximality. Thus $o'_{x_3} = o''_{x_3}$. Together with $o_{x_3} > o'_{x_3}$, this implies P_1.

Proof of $\forall y, P_2(y)$. Consider a modality $y_0 \in \mathcal{M}$. If $o_{y_0} \geq o''_{y_0}$, $P_2(y_0)$ is trivially satisfied. Otherwise, we have $o_{y_0} < o''_{y_0}$ (H_1). We then have two cases :

(A) Suppose $o_{y_0} < o'_{y_0}$ (H_2). By Lemma 1, E_2 and H_2, $S_2 = \{x \in \mathcal{M} | x \succ_m y_0 \wedge o_x > o'_x\}$ is not empty, so we can choose x' in $max_{\succ_m} S_2$.

(A.1) Suppose $o'_{x'} < o''_{x'}$ (H_3). By F_2 and H_3, we get a modality z such that $z \succ_m x' \wedge o'_z > o''_z$. We have $z \succ_m x'$ and $x' \succ_m y_0$, thus, by transitivity of $\succ_m$, $z \succ_m y_0$. Since this and the fact that x' is maximal for $\succ_m$, we must have $z \notin S_2$ which gives us $o_z \geq o'_z$. Having $o_z > o'_z$ is not possible as it would allow us to derive from E_2 a modality that would belong to S_1 while being superior to x', contradicting again the maximality of x'. We can conclude $o_z = o'_z$, and thus $o_z > o''_z$, which proves $P_2(y_0)$.

(A.2) Otherwise, $o'_{x'} \geq o''_{x'}$. Given $x' \in S_1$, this implies $o_{x'} > o''_{x'}$. We thus also have (taking x' for z), $P_2(y_0)$.

(B) Otherwise, $o_{y_0} \geq o'_{y_0}$ (H_4). We then consider the modalities that are superior to y_0.

(B.1) Suppose that $\exists y' \in \mathcal{M}, y' \succ_m y_0 \wedge o_{y'} < o'_{y'}$. Then, by applying reasoning of case A.1 to y', we get some $z \in \mathcal{M}$ such that $z \succ_m y'$ and $o_z > o''_z$. By transitivity of $\succ_m$, $z \succ_m y_0$, which proves $P_2(y_0)$.

(B.2) Otherwise, we must have : $\forall y' \in \mathcal{M}, y' \succ_m y_0 \Rightarrow o_{y'} \geq o'_{y'}$ (H_5). From H_1 and H_4, we have $o'_{y_0} < o''_{y_0}$. Applying F_2 gives a modality z such that $z \succ y_0$ and $o'_z > o''_z$. Given H_5 we get $o_z \geq o'_z$ and thus $o_z > o''_z$, which proves again $P_2(y_0)$.

We have thus proven $P_2(y_0)$ in all cases for any y_0.

This concludes the proof of the theorem 1. $\succ_e^m$ is indeed a strict partial order.

5 Short Discussion on the Assumptions Made on $\succ_m$

It has been assumed that the $\succ_m$ relation is asymmetric and transitive, it already encompasses many situations, but this is not necessarily the case. We briefly discuss in this section two alternatives.

5.1 Case of a Connected $\succ_m$

Adding other assumptions can give additional information about $\succ_e$. For example, if we assume that $\succ_m$ is also connected, then our axiomatisation becomes a lexicographic ordering on the modalities [5]. Thus for any non-equal pair of actions o and o', the proposed axiomatisation in Eq. 8 will infer a preference. This property is useful if we want an unique action to perform. However if only one action is permissible, it is a restrictive property for an ethical compliance system.

5.2 Case of a Non Transitive $\succ_m$

One can also wish that the relation $\succ_m$ is not transitive. However this is not a very credible case in an ethical decision context where we define superiorities between modalities. Doing so would allow loops: x is superior to y which is superior to z which is superior to x. Nevertheless, removing this hypothesis implies a modification of the axiomatisation. Indeed the transitivity of $\succ_m$ is essential to the proofs of asymmetry and transitivity for $\succ_e$.

6 Conclusion and Future Works

This paper proposes an axiomatisation of the philosophical concept of superiority between the modalities of the Good. To do so, an ordinal multi-criteria decision formalism adapted to ethical decision making has been defined, based on a utilitarian approach. As a generalisation of the Pareto optimality principle, the proposed axiomatisation makes it possible to deduce preferences from the differentiation of modalities.

As this paper is a first attempt to connect philosophical concerns with ordinal preferences, ongoing works aim at studying existing formal frameworks that offer properties similar to the ones we propose, such as e.g. hard and soft constraints hierarchies [2].

The work presented in this paper also opens multiples perspectives. Firstly, one limitation of the current work lies in the simplifications made on the causal relations in our formalism. In a more concrete problem, it would be ethically more precise to take into account each consequence separately. Therefore, we consider to extend the formalism to be able to encompass such cases.

Furthermore, the ability to extend the minimal ethical preference set that respect superiority raises questions about mixing multiples principles to get a single set of permissible actions. Thus, we intend to formalise a more generalised version of the concept of ethical principle and the constraints that they must fulfill.

References

1. Anderson, M., Anderson, S.L.: Machine Ethics. Cambridge University Press (2011). https://doi.org/10.1017/CBO9780511978036
2. Borning, A., Freeman-Benson, B., Wilson, M.: Constraint hierarchies. LISP and symbolic computation **5**(3), 223–270 (1992)
3. Bourgne, G., Sarmiento, C., Ganascia, J.G.: Ace modular framework for computational ethics: dealing with multiple actions, concurrency and omission. In: 1st International Workshop on Computational Machine Ethics, Online event (2021)
4. Chang, R.: Incommensurability (and Incomparability). John Wiley & Sons, Ltd (2013). https://doi.org/10.1002/9781444367072.wbiee030
5. Fishburn, P.C.: Axioms for Lexicographic Preferences. The Review of Economic Studies **42**(3), 415–419 (07 1975). https://doi.org/10.2307/2296854
6. Griffin, J.: Are there incommensurable values? Philosophy & Public Affairs **7**(1), 39–59 (1977), http://www.jstor.org/stable/2265123
7. Lindner, F., Bentzen, M.M., Nebel, B.: The hera approach to morally competent robots. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6991–6997. IEEE (2017)
8. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: A survey. ACM Comput. Surv. **53**(6) (12 2021). https://doi.org/10.1145/3419633
9. Vallentyne, P.: Consequentialism. In: Philosophy publications. Wiley-Blackwell (2006)