Information Sciences 631 (2023) 288–304

Contents lists available at ScienceDirect

Information Sciences

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

# Building rankings encompassing multiple criteria to support qualitative decision-making

Marc Serramia [a,*], Maite Lopez-Sanchez [b], Stefano Moretti [c], Juan A. Rodriguez-Aguilar [d]

[a] King's College London, 30 Aldwych, WC2B 4BG, London, United Kingdom
[b] University of Barcelona, Gran Via de les Corts Catalanes, 585, 08007, Barcelona, Spain
[c] Laboratoire d'Analyse et Modélisation de Systèmes pour l'Aide à la décision (LAMSADE), Place du Maréchal de Lattre de Tassigny, 75016, Paris, France
[d] Artificial Intelligence Research Institute (IIIA-CSIC), Campus UAB, Carrer de Can Planas, Zona 2, 08193, Bellaterra, Spain

ARTICLE INFO

*Keywords:*
Decision support
Computational social choice
Ethical decision making

ABSTRACT

Decision makers are commonly challenged with comparing, and ultimately ranking, elements with regards to the degree to which they satisfy multiple criteria and in terms of their own preferences. This calls for a new decision making framework, which we formally present here. Within such a framework, we present *multi-criteria lex-cel*: a new method for ranking single elements. Furthermore, we formally establish that our contributions generalise recent results in the social choice literature. We also illustrate our contributions through a case study that poses an *ethical* decision-making problem.

## 1. Introduction

Rankings establish comparisons between individual objects (or sets of objects) that are useful for many applications. Consider, for example, the widely studied problem of college admissions [49]. Ranking solutions have been proposed to solve this problem [39] and other similar problems such as committee selection [21]. Thus, they have been long investigated in the literature. Without aiming for completeness, here we highlight three different bodies of work related to rankings. Firstly, the literature has countless examples of works studying voting and ranking aggregation. A representative example of recent developments in this area includes the work of Aledo et al. [2] on a highly scalable algorithm to aggregate general rankings, and Miebs et al. [30] who study heuristic algorithms to aggregate incomplete partial rankings. Secondly, Barbera et al. [8] study functions that transform rankings of individual elements into rankings of sets of these elements. Maxmin and minmax [5] or leximin and leximax [38] are examples of such functions. Thirdly, Moretti and Ozturk [35] introduce the social ranking as a mapping that transforms a ranking of sets of elements into a ranking of the individual elements of these sets. Social rankings have been extensively studied: Haret et al. [19] base their work on the *ceteris paribus* majority principle; Khani et al. [20] focus on the notion of marginal contribution; and Doignon et al. [14] study the stability of social scorings (a concept related to social rankings). We can even find the usage of social rankings in ethical decision-making [43].

* Corresponding author.
*E-mail address:* marc.serramia_amoros@kcl.ac.uk (M. Serramia).

https://doi.org/10.1016/j.ins.2023.02.063
Received 20 May 2021; Received in revised form 15 February 2023; Accepted 18 February 2023

Available online 24 February 2023

A particularly interesting social ranking is lex-cel introduced by Bernardi et al. [12] which focuses on lexicographical preferences and satisfies some desirable properties. Lex-cel has caught the attention of the social choice community, so much so that there have recently been many works, generalising it [3,9], proving it is not manipulable [4], applying it to coalition formation [22], and defining a new social choice function based on it [27].

A common assumption in the ranking literature is the existence of preferences regarding the elements or sets of elements from which to build a ranking. Indeed, this assumption is reasonable when considering a limited number of candidate elements — e.g. in a presidential election. However, when considering many candidates – e.g. in scholarship assignments – it is common practice simply to be provided with the criteria for establishing the element ranking, rather than an explicit ranking over the individual elements. Moreover, if multiple criteria are considered, we may also be provided with preferences regarding them. For example, when designing a diet, we may consider different criteria regarding the food to choose – such as its healthiness or tastiness – and preferences over these criteria — e.g., healthiness preferred over tastiness.

Against this background, in this paper we develop novel formal tools to help decision-makers take qualitative decisions about multiple options while considering their preferences. In particular, we propose *multi-criteria-based ranking* (MC ranking), a method to rank individual elements based on: i) how they relate to the criteria (e.g., if the food is tasty); and ii) the preferences over these criteria. Specifically, when building MC rankings, we consider that candidate elements may relate to a given criterion to different degrees. Thus, following the example of a diet, broccoli cheddar soup can be considered to be not particularly healthy, a Caesar salad to be healthy, and steamed vegetables to be very healthy. Moreover, these qualitative relationships may even be negative — with sausages considered unhealthy and chips very unhealthy. Overall, our multi-criteria-based ranking encompasses rich qualitative element-criterion relations that produce a comprehensive ranking over the individual candidate elements. Briefly, the contributions of this paper are:

1. A formal definition of a new type of rankings: *multi-criteria-based rankings* (MC rankings).
2. A formal definition of dominance for MC rankings. This definition requires a non-straightforward adaptation from the desirable dominance property in social choice.
3. A definition and study of the so-called *multi-criteria lex-cel*, a function to create MC rankings embodying dominance.
4. A formal analysis showing the generality of our contributions with respect to recent results in the literature. Interestingly, *MC rankings* generalise social rankings [35], while *multi-criteria lex-cel* generalises the lex-cel ranking function introduced in [12].
5. A case study posing an ethical decision-making problem that illustrates the use of MC rankings.

This paper is organised as follows. We first introduce the necessary background to order theory in Section 2, while Section 3 formalises labels and label systems. Next, Section 4 formalises MC rankings, as well as the property of dominance, while Section 5 introduces MC lex-cel. Subsequently, Section 6 studies the relation of our MC rankings and MC lex-cel with the literature on social choice. Finally, Section 7 analyses a case study in ethical decision-making, while Section 8 shows an application in participatory budgets, and Section 9 discusses our conclusions.

## 2. Background: order theory

Let $X$ be a set of objects. A binary relation $\succeq$ on $X$ is said to be: *reflexive*, if for each $x \in X$, $x \succeq x$; *transitive*, if for each $x, y, z \in X$, $(x \succeq y$ and $y \succeq z) \Rightarrow x \succeq z$; *total*, if for each $x, y \in X$, $x \succeq y$ or $y \succeq x$; *antisymmetric*, if for each $x, y \in X$, $x \succeq y$ and $y \succeq x \Rightarrow x = y$. We can define preferences among the elements of $X$ by means of binary relations. Moreover, we can categorise the type of preferences depending on the properties they hold as follows.

**Definition 1** *(Preorder, ranking, linear order and partial order).* A *preorder* (or *quasi-ordering*) is a binary relation $\succeq$ that is reflexive and transitive. A preorder that is also total is called a *total preorder* or *ranking*. A total preorder that is also antisymmetric is called a *linear order*. A preorder that is antisymmetric but not total is called *partial order*.

We build a lexicographical order for two tuples by comparing them element-wise from left to right. While the elements in both tuples are the same, we move to the next position on the tuples. We traverse the tuples until two elements differ (one is preferred over the other). The more preferred tuple is the one containing the more preferred element. If all elements are the same, the tuples are deemed equal. Formally:

**Definition 2.** Given two tuples $t, t'$, with $t = (t_1, \dots t_q)$ and $t' = (t'_1, \dots t'_q)$, we define the lexicographical order of tuples $\succeq_{lex}$ as: $t \succeq_{lex} t' \Leftrightarrow$ if either $t = t'$ or $\exists i \in \{1, \dots q\}$ s.t. $t_i > t'_i$ and $\forall j < i, t_j = t'_j$ (note that $t =_{lex} t' \Leftrightarrow t = t'$).

The lexicographical order for tuples is used in the definition of the lex-cel ranking [12]. Let $X$ be a set of elements, and $\succeq_S$ a ranking over the power set $\mathcal{P}(X)$, then lex-cel builds an element ranking $\succeq_e$ by means of assigning a tuple to each element (noted $\theta(x)$). To build this tuple, consider the quotient set $\mathcal{P}(X)/\sim_S$ with quotient order $\Sigma_1 \succ_S \Sigma_2 \succ_S \cdots \succ_S \Sigma_q$. Then, $\theta(x)$ is defined as:

$$\theta(x) = (x_1, \dots, x_q) \text{ where } x_i = |S \in \Sigma_i : x \in S| \tag{1}$$

Lex-cel ranks elements in $X$ by comparing lexicographically their corresponding $\theta$ tuples: $x \succeq_e y \Leftrightarrow \theta(x) \succeq_{lex} \theta(y)$.

## 3. Relating elements to criteria

As previously introduced, when ranking candidate elements according to given criteria, we consider those candidate elements that are related to the given criteria. Specifically, we enrich the expressivity of these relations by means of graded labels. As we assume humans will assign semantics to the labels and will specify these relations, we take inspiration from the widely-used Likert scale [36],[1] and specify that graduation ranges from negative labels, which signal that an element is detrimental to a criterion, through a neutral label, to positive labels, which indicate that an element aligns with a criterion. Next, we introduce the notions of label system, the object that defines labels for relating elements and criteria and their semantics, and labelling, a function to relate elements to criteria through labels.

A label system contains a set of labels and an order over them to establish their grading. This set of labels must contain a neutral label, lying between positive and negative labels. Positive labels are those that are more preferred than the neutral label, whereas negative labels are those that are less preferred than the neutral label. In terms of label grading, the more preferred a positive label, the higher the degree of alignment between an element and the criterion it is meant to represent. Conversely, the less preferred a negative label, the higher the detrimental degree.

**Definition 3** *(Label system).* A label system is a pair $\langle L, >_L \rangle$, where $L$ is a set of labels, and $>_L$ is a linear order over $L$. A label system includes a neutral label[2] $l_0 \in L$. Labels more preferred than $l_0$ are positive labels, whereas those less preferred than $l_0$ are negative labels.

Note that a label system does not need to have a negative label for each positive label. In fact, it might only have positive labels. However, a label system with more labels of one type than of another one hinders the task of comparing labels. For example, given $l_2 >_L l_1 >_L l_0 >_L l_{-1}$, it is unclear whether the positive counterpart of $l_{-1}$ is $l_2$ because both labels are the most extreme ones, or if it is $l_1$, because they are equally separated from $l_0$. To avoid these uncertainties, we focus on a particular type of label systems: the so-called symmetric label systems, for which each positive label has a negative counterpart. To ease their definition, we first introduce two auxiliary functions, namely the sign and strength of a label, which also provide a useful notation for the forthcoming sections.

Given a label system, the sign function signals whether a label is positive (1), negative (-1), or the neutral label (0).

$$sgn(l) = \begin{cases} 1 & \text{if } l >_L l_0 \\ 0 & \text{if } l = l_0 \\ -1 & \text{if } l_0 >_L l \end{cases} \tag{2}$$

The strength function characterises the label's degree of preference in the label system order. In particular, we consider that, given a label $l$, the more labels between $l$ and $l_0$ in the label order, the greater its strength. Formally:

$$stg(l) = \begin{cases} |\{l' \in L, l \geq_L l' >_L l_0\}| & \text{if } l >_L l_0 \\ 0 & \text{if } l = l_0 \\ |\{l' \in L, l_0 >_L l' \geq_L l\}| & \text{if } l_0 >_L l \end{cases} \tag{3}$$

**Definition 4** *(Symmetric label system).* A label system $\langle L, >_L \rangle$ is symmetric if $\forall l \in L$, $\exists l' \in L$, such that $sgn(l) = -sgn(l')$ and $stg(l) = stg(l')$.

Symmetric label systems have the same number of positive and negative labels. Note that, without loss of generality, any label system can be transformed into a symmetric label system by simply adding superfluous labels. Hereafter, we only consider symmetric label systems. Also, we can uniquely note each label in the label system as $l_{sgn(l)\cdot stg(l)}$ (for example, we note as $l_{-2}$ the label of sign -1 and strength 2).

**Example 1.** Consider $L' = \{l_1, l_0, l_{-1}, l_{-2}\}$ to be a label system with order $l_1 >_L l_0 >_L l_{-1} >_L l_{-2}$. Note, for example, that $l_{-2}$ is the label of sign $sgn(l_{-2}) = -1$ and strength $str(l_{-2}) = 2$. Moreover, as we require symmetry, we can add an additional superfluous label $l_2$ to transform the system into a symmetric label system $L = \{l_2, l_1, l_0, l_{-1}, l_{-2}\}$ with order $l_2 >_L l_1 >_L l_0 >_L l_{-1} >_L l_{-2}$.

Using a label system, a decision-maker can relate an element with a criterion by means of a labelling function.

**Definition 5.** Given a set of elements $X$, a set of criteria $C$, and a label system $\langle L, >_L \rangle$, a labelling is a function $\lambda : X \times C \to L$ that assigns a label in $L$ to each pair of elements in $X$ and criterion in $C$, thereby establishing the relation between the element and the criterion. We note as $\mathcal{L}(X, C)$ the set of all possible labellings over $X$ and $C$.

[1] The Likert scale is a psychometric scale [36] ubiquitous in survey research. It is recognised as universally applicable as attitudes towards any object or on any issue can vary along the same underlying negative-to-positive dimension.

[2] $l_0$ is unique because $>_L$ is a linear order. Thus, if there were two neutral labels, one would be necessarily preferred over the other.

If $\lambda(x,c)=l$, we say that element $x$ is related to criterion $c$ with degree $l$. From equation (3), we also say that the strength of the relation is $stg(l)$. For example, recalling the diet example, a labelling would relate steamed vegetables ($sv$) to the healthiness criterion ($h$) with a label of positive sign and very high strength ($l_{max}$), thus resulting in $\lambda(sv,h)=l_{max}$.

**Example 2.** Consider the set of elements $X=\{x_1,\dots,x_5\}$ and a set of criteria $C=\{c_1,\dots,c_4\}$ and the label system of Example 1. An example of a labelling would be:

| $\lambda$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $c_1$ | $l_{-2}$ | $l_1$ | $l_1$ | $l_0$ | $l_0$ |
| $c_2$ | $l_2$ | $l_1$ | $l_0$ | $l_0$ | $l_0$ |
| $c_3$ | $l_0$ | $l_{-1}$ | $l_0$ | $l_{-1}$ | $l_2$ |
| $c_4$ | $l_0$ | $l_{-1}$ | $l_0$ | $l_{-2}$ | $l_0$ |

## 4. Multi-criteria-based rankings

As mentioned above, we assume that the decision maker establishes a set of criteria and knows their preferences over them. We have learnt in Section 3 how to relate elements to criteria. Our goal is to build a ranking of the single elements in $X$ from: (i) the relationships between elements and criteria; and (ii) preferences regarding criteria. We will call such ranking a *multi-criteria-based ranking* (MC ranking). In this section, we formally define this ranking, as well as the fundamental notion of dominance for MC rankings.

An MC ranking considers a set of elements $X$, a set of criteria $C$, a ranking $\succeq_C$ over the criteria, and a labelling $\lambda$ relating elements to criteria and builds a ranking $\succeq$ over the single elements in $X$. Formally:

**Definition 6.** Given a set of elements $X$, a set of criteria $C$, and a set of labellings $\mathcal{L}(X,C)$, an MC ranking is a function $mcr : \mathcal{L}(X,C)\times\mathcal{R}(C)\to\mathcal{R}(X)$ that associates any pair of labelling $\lambda\in\mathcal{L}(X,C)$ (relating elements with criteria) and ranking $\succeq_C\in\mathcal{R}(C)$ (over the criteria) to another ranking $mcr(\lambda,\succeq_C)\in\mathcal{R}(X)$ over the elements of $X$.

MC rankings call for the introduction of a novel notion of dominance between the elements in $X$, as is common in the literature (e.g. [8] [35]). Such notion of dominance must ensure that the ranking of elements is based strictly on the ranking over criteria. However, defining dominance for MC rankings is intricate due to the richness of our labelling approach. Informally, our notion of dominance requires that an MC ranking function ranks the elements in $X$ taking into accountthe element-criterion relations, their associated labels, and the criteria preferences. Thus, the more preferred a criterion with which an element relates positively, the more preferred the element. Conversely, the more preferred the criterion with which an element relates negatively, the less preferred the element. The higher the degree of the labels on these positive/negative relations, the more/less preferred the element will be. Furthermore, the larger the number of positive relations and the lower the number of negative relations for an element, the more preferred the element in the ranking will be.

Our notion of dominance between two elements is founded on the dominance within each equivalence class of criteria resulting from the ranking $\succeq_C$ over criteria. Thus, consider the quotient set of criteria $C/\sim_C$ with equivalence classes $\kappa_1,\dots,\kappa_r$, and quotient order $\succ_C$. Note that the criteria within each equivalence class $\kappa\in C/\sim_C$ are preferred equally. Given an equivalence class $\kappa$, our first aim is to establish whether an element $x\in X$ is $\kappa$-dominant (dominant within the scope of the equivalence class $\kappa$) over another element $y\in X$. An element will be $\kappa$-dominant over another if it relates more strongly (and positively) to the criteria in $\kappa$ than another element.

To define $\kappa$-dominance, we employ an auxiliary function, the so-called net alignment function. Given an element $x$ and a strength $s$, the net alignment function aggregates the positive and negative relations of $x$ with the criteria in $\kappa$ with strength $s$. Thus, the larger the net alignment, the more positive relations of strength $s$ relating $x$ and $\kappa$, and the lower the net alignment, the more negative relations of strength $s$ relating $x$ and $\kappa$. Formally, the net alignment function (noted $na$) is defined as the difference between the number of criteria positively and negatively related to $x$ with strength $s$:

**Definition 7.** Consider a criteria equivalence class $\kappa\in C/\sim_C$ and a relation strength $s\neq 0$. We define the *net alignment* of strength $s$ of element $x$ with class $\kappa$ as:

$$na(x,\kappa,s)=|\{c\in\kappa:\lambda(x,c)=l_s\}|-|\{c\in\kappa:\lambda(x,c)=l_{-s}\}| \tag{4}$$

Let $s_{max}=\max_{l\in L} stg(l)$ be the maximum strength of the labels in the label system. Then, $\kappa$-dominance is defined as:

**Definition 8.** Given two elements $x,y\in X$, a set of criteria $C$, a ranking over these criteria $\succeq_C$, a symmetric label system $\langle L,>_L\rangle$ and a criteria equivalence class $\kappa\in C/\sim_C$, we say that $x$ is $\kappa$-dominant over $y$ if $\exists s\in\{1,\dots,s_{max}\}$, s.t. $na(x,\kappa,s)>na(y,\kappa,s)$ and $\forall s'>s$, we have $na(x,\kappa,s')=na(y,\kappa,s')$. If $\forall s\in\{1,\dots,s_{max}\}$, $na(x,\kappa,s)=na(y,\kappa,s)$, we say $x$ and $y$ are $\kappa$-indifferent.

**Example 3.** Following Example 2, consider the criteria preferences $c_1 \succeq_C c_2 \sim_C c_3 \sim_C c_4$. The quotient set is $C/\sim_C = \{\kappa_1, \kappa_2\}$, with $\kappa_1 = \{c_1\}$ and $\kappa_2 = \{c_2, c_3, c_4\}$, and quotient order $\kappa_1 \succ_C \kappa_2$. Note that $na(x_1, \kappa_1, 2) = -1$, while for the rest of the elements in $X$, their net alignment of strength 2 with $\kappa_1$ is 0, which is greater than $-1$. Thus, we say that $x_2, x_3, x_4$, and $x_5$ are $\kappa_1$-dominant over $x_1$.

Notice that the relation over the elements in $X$ provided by $\kappa$-dominance corresponds to a lexicographical order over the tuples of net alignment values arranged in decreasing order of label strength, from $s_{max}$ to 1 (see relation (6) in the following section for a formal definition). In the previous example, with $s_{max} = 2$, the tuples of net alignment values $(na(x, \kappa_2, 2), na(x, \kappa_2, 1))$ on the criteria equivalence class $\kappa_2$ for each element in $x \in X$ are $(1, 0)$ for $x_1$ and $x_5$, $(0, -1)$ for $x_2$, $(0, 0)$ for $x_3$ and $(-1, -1)$ for $x_4$. Therefore, the lexicographical order induced by $\kappa_2$-dominance on the tuples of net alignment values is $(1, 0) >_{lex} (0, 0) >_{lex} (0, -1) >_{lex} (-1, -1)$. This corresponds to stating that $x_1$ and $x_5$ are $\kappa_2$-indifferent and $\kappa_2$-dominant over the other elements, $x_3$ is $\kappa_2$-dominant over $x_2$ and $x_4$ and, finally, $x_2$ is $\kappa_2$-dominant over $x_4$. The reason for assuming lexicographical orders over the tuples of net alignments rests on a principle aimed at preventing any kind of trade-off or compensation between labels of different strengths (for instance, a relation of the highest grade between an element and a single criterion cannot be balanced by many relations of moderate degree between the element and other criteria in the same equivalence class). In other words, the choice of a lexicographical order over tuples of net alignment values boils down to defining a rigorous hierarchy over the labels to favour the excellence of elements over criteria in the same equivalence class. Although the choice of such a hierarchy over labels depends on the application context, the use of lexicographical orders in models of preference and choice has been widely analysed in the literature of decision theory and artificial intelligence, and its application has been studied extensively both from a mathematical and a pragmatic perspective (see for instance [16,23,42]).

Using the concept of $\kappa$-dominance, we define dominance considering all equivalence classes in $C/\sim_C$ (and their quotient order $\succ_C$). We say that $x$ is dominant over $y$ if for a given criteria equivalence class $x$ is $\kappa$-dominant over $y$, while for more preferred equivalence classes they are $\kappa$-indifferent.

**Definition 9.** Given two elements $x, y \in X$ with criteria in $C$ and a ranking over criteria $\succeq_C$, we say that $x$ is dominant over $y$ if there is a criteria equivalence class $\kappa \in C/\sim_C$, such that: (i) $x$ is $\kappa$-dominant over $y$; and (ii) $\forall \kappa' \in C/\sim_C$, such that $\kappa' \succ_C \kappa$, $x$ and $y$ are $\kappa'$-indifferent. If neither element dominates the other (they are $\kappa$-indifferent $\forall \kappa \in C/\sim_C$), we say that they are indifferent.

Dominance is a natural extension of the $\kappa$-dominance notion to a multi-criteria framework where a preference ranking $\succeq_C$ over criteria is given. As we will explain in detail in Section 5.2, the dominance relation over elements represent a lexicographical order aimed at rewarding the elements having excellent labelling degrees in the most preferred criteria equivalence classes.

**Example 4.** From $\kappa$-dominance in Example 2, we conclude that $x_2, x_3, x_4$ and $x_5$ are dominant over $x_1$ because they are $\kappa_1$-dominant and that $\kappa_1$ is the most preferred class.

## 5. Multi-criteria lex-cel

Next, we introduce multi-criteria lex-cel (MC lex-cel), which is an MC ranking function. For each element in $X$, MC lex-cel builds a tuple, the so-called *multi-criteria profile* (MC profile), which summarises the relations between the element and the criteria. Then, MC lex-cel ranks the elements in $X$ by comparing their MC profiles lexicographically. In Section 5.1, we describe how to build MC profiles, whereas Section 5.2 defines MC lex-cel and proves that it embodies the dominance property in Definition 9.

### 5.1. Building MC profiles for elements

We will build the MC profile of an element $x \in X$ as a tuple $\mu(x)$ that is meant to summarise the relations of that particular element with all the criteria at hand.

In general terms, we build an MC profile for an element through a nested process: (1) we start considering criteria preferences, from more preferred to less preferred; (2) thereafter, we delve into each equivalence class to consider the strengths of the relations, from stronger to weaker.

Formally, we build an MC profile by considering the quotient set $C/\sim_C$, where $\kappa_1, \dots, \kappa_q \in C/\sim_C$ are criteria equivalence classes with quotient order $\kappa_1 \succ_C \cdots \succ_C \kappa_q$. Each $\kappa_i$ contains the $i$-th most preferred criteria.

We compose the MC profile $\mu(x)$ of an element $x$, from its equivalence class profiles $\mu(x, \kappa_1), \dots, \mu(x, \kappa_q)$. An equivalence class profile $\mu(x, \kappa_i)$ summarises the relations between $x$ and the equivalence class $\kappa_i$. We want to ensure that criteria preferences are satisfied according to $\succ_C$. Thus, we compose the MC profile $\mu(x)$ by considering that the relationships with more preferred criteria are positioned further to the left[3] of $\mu(x)$ as follows:

$$\mu(x) = (\mu(x, \kappa_1), \dots, \mu(x, \kappa_q)) \tag{5}$$

[3] Recall that the MC lex-cel function in Section 5.2 applies a lexicographical order over $\mu(x)$, and thus the left indicates greater preference.

Within an equivalence class $\kappa$, all criteria are preferred indifferently. Thus, what distinguishes the relations between $x$ and $\kappa$ here is their strength and sign. Recall that for each strength $s$, the net alignment function $na$ aggregates the number of positive relations of strength $s$ with the number of negative relations of strength $s$. Hence, we build the equivalence class profile of $x$ for class $\kappa$ out of the net alignments between $x$ and $\kappa$ for all non-zero[4] strengths, namely from $na(x,\kappa,1),\dots,na(x,\kappa,s_{max})$. Since we prefer strong relations over weak ones, the net alignments representing greater strengths, are positioned further to the left[3]. Therefore, the equivalence class profile is a tuple containing the net alignments of $x$ and $\kappa$ arranged from left to right in descending order of strength:

$$\mu(x,\kappa)=(na(x,\kappa,s_{max}),\dots,na(x,\kappa,1)), \tag{6}$$

where, as for Definition 8, $s_{max}=\max_{l\in L} stg(l)$ is the maximum strength of the label system. For the sake of understanding, we will now illustrate how to build the MC profiles for the elements in our running example.

**Example 5.** Following our running example, note that the criteria preferences $c_1 \succeq_C c_2 \sim_C c_3 \sim_C c_4$ imply that $C/\sim_C = \{\kappa_1,\kappa_2\}$, with $\kappa_1=\{c_1\}$, $\kappa_2=\{c_2,c_3,c_4\}$, and $\kappa_1 \succ_C \kappa_2$. Thus, $\forall x\in X$, $\mu(x)=(\mu(x,\kappa_1),\mu(x,\kappa_2))$. Now, the label system that we have considered contains labels of strength 2, 1 (and 0). Hence, since the maximum strength is 2, $\mu(x,\kappa)=(na(x,\kappa,2),na(x,\kappa,1))$ for each element $x$. In particular, regarding $x_1$, we have that $na(x_1,\kappa_1,2)=-1$, because there is one label $l_{-2}$ relating $x_1$ to $\kappa_1$, whereas $na(x_1,\kappa_1,1)=0$, because there are no labels of strength 1 relating $x_1$ to $\kappa_1$. By applying equation (6) above, we have that $\mu(x_1,\kappa_1)=(-1,0)$. On the other hand, we have it that $na(x_1,\kappa_2,2)=1$ because there is one label $l_2$ relating $x_1$ to $\kappa_2$, while $na(x_1,\kappa_2,1)=0$, because there are no labels of strength 1 relating $x_1$ to $\kappa_2$. Again, by means of equation (6), we have that $\mu(x_1,\kappa_2)=(1,0)$. With these two equivalence class profiles, we can now apply equation (5) to build the MC profile of $x_1$ as $\mu(x_1)=((-1,0),(1,0))$. By following an analogous procedure, we obtain the MC profiles for the rest of elements of $X$:

$$\mu(x_2)=((0,1),(0,-1)) \qquad \mu(x_3)=((0,1),(0,0))$$
$$\mu(x_4)=((0,0),(-1,-1)) \qquad \mu(x_5)=((0,0),(1,0))$$

### 5.2. The multi-criteria lex-cel ranking function

Since the MC profile of an element $x\in X$ encodes its alignment with the criteria in $C$, we propose comparing elements in $X$ by comparing their MC profiles by means of their lexicographical order. This is precisely what our multi-criteria lex-cel function captures as follows:

$$x \gtrsim y \Leftrightarrow \mu(x)\geq_{lex}\mu(y).$$

**Definition 10.** Given a set of elements $X$, a set of criteria $C$ and a set of labellings $\mathcal{L}(X,C)$, the multi-criteria lex-cel (MC lex-cel) function $mclex:\mathcal{L}(X,C)\times\mathcal{R}(C)\to\mathcal{R}(X)$ associates to any labelling $\lambda\in\mathcal{L}(X,C)$ and any ranking $\succeq_C\in\mathcal{R}(C)$, another ranking $\gtrsim = mclex(\lambda,\succeq_C)\in\mathcal{R}(X)$ such that for any two elements $x,y\in X$:

$$x \gtrsim y \Leftrightarrow \mu(x)\geq_{lex}\mu(y), \tag{7}$$

where $>_{lex}$ the lexicographical order in Definition 2.

Notice that $\mu(x)>_{lex}\mu(y)\Leftrightarrow\exists\kappa\in C/\sim_C$, such that $\forall\kappa'\succ_c\kappa\ \mu(x,\kappa')=\mu(y,\kappa')$ and $\mu(x,\kappa)>_{lex}\mu(y,\kappa)$.

**Example 6.** After applying MC lex-cel to the MC profiles obtained in Example 5, we obtain the following element ranking: $x_3\succ x_2\succ x_5\succ x_4\succ x_1$.

Our purpose now is to prove that MC lex-cel embodies dominance according to Definition 9. Before doing that, we need an intermediary result showing that the lexicographical ordering of criteria profile captures $\kappa$-dominance within criteria equivalence classes.

**Lemma 1.** *Consider two elements $x,y\in X$, and a criteria equivalence class $\kappa\in C/\sim_C$, then $\mu(x,\kappa)>_{lex}\mu(y,\kappa)\Leftrightarrow x$ $\kappa$-dominant over $y$. Otherwise, we have that $\mu(x,\kappa)=\mu(y,\kappa)\Leftrightarrow x$ and $y$ are $\kappa$-indifferent.*

**Proof.** Suppose that $\mu(x,\kappa)>_{lex}\mu(y,\kappa)$. Since $\mu(x,\kappa)=(na(x,\kappa,s_{max}),\dots,na(x,\kappa,1))$ and $\mu(y,\kappa)=(na(y,\kappa,s_{max}),\dots,na(y,\kappa,1))$, with $\mu(x,\kappa)$ being lexicographically greater than $\mu(y,\kappa)$ means that $\exists s\in\{1,\dots,s_{max}\}$, such that $na(x,\kappa,s)>na(y,\kappa,s)$, and $\forall s'>s$, $na(x,\kappa,s')=na(y,\kappa,s')$. Notice that this is precisely the definition of $x$ $\kappa$-dominant over $y$ (Definition 8). Now, if $\mu(x,\kappa)=\mu(y,\kappa)$, then $\forall s\in\{1,\dots,s_{max}\}$, $na(x,\kappa,s)=na(y,\kappa,s)$, which is the definition of $x$ and $y$ being $\kappa$-indifferent. Consider now the other direction of the implication, and then suppose that $x$ is $\kappa$-dominant over $y$. In this case, neither $\mu(y,\kappa)>_{lex}\mu(x,\kappa)$ nor $\mu(x,\kappa)=\mu(y,\kappa)$ can

[4] A strength zero relation (labelled $l_0$) represents that the element is neutral to the criterion. In other words, the element does not affect the criterion (the element neither aligns with nor is detrimental to the criterion). Hence, we should not take into account these relations in the MC profile.

be true because, by the already proved implication, it would contradict our assumption. Therefore, the only possibility is that $\mu(x,\kappa) >_{lex} \mu(y,\kappa)$. The same reasoning applies if we suppose $y$ is $\kappa$-dominant over $x$, or $x$ and $y$ are $\kappa$-indifferent.

With the help of Lemma 1, we are now ready to prove that multi-criteria lex-cel embodies dominance.

**Theorem 1.** *MC lex-cel embodies dominance, that is, if* $mclex(\succeq_C) = \succeq$, *then for* $x, y \in X$, *we have that* $x \succ y \Leftrightarrow x$ *is dominant over* $y$.

**Proof.** Suppose that $x \succ y$. Since $\succ$ has been obtained through MC lex-cel, we know that $\mu(x) >_{lex} \mu(y)$. This means that $\exists \kappa \in C/\sim_C$, such that $\mu(x,\kappa) >_{lex} \mu(y,\kappa)$ and $\forall \kappa' \succ_C \kappa$, $\mu(x,\kappa') = \mu(y,\kappa')$. Thanks to Lemma 1, we have seen that this means that $x$ is $\kappa$-dominant over $y$ and $\forall \kappa' \succ_C \kappa$, $x$ and $y$ are $\kappa'$-indifferent, which is the definition of dominance of $x$ over $y$. Similarly, if $\mu(x) = \mu(y)$, then $\forall \kappa$, $\mu(x,\kappa) = \mu(y,\kappa)$, and thus $x$ and $y$ are $\kappa$-indifferent, meaning that they are indifferent. As to the other direction of the proof, say that $x$ is dominant over $y$. If $\mu(x) <_{lex} \mu(y)$, it would imply that $y$ is dominant over $x$, which contradicts our assumption. Similarly, if $\mu(x) = \mu(y)$, $x$ and $y$ should be indifferent, again contradicting our assumption. Therefore, the only possibility is that $\mu(x) >_{lex} \mu(y)$. The same reasoning applies if we suppose that $y$ is dominant over $x$ or $x$ and $y$ are indifferent.

## 6. MC ranking and social ranking

In this section we explore the relation between our MC ranking and the social ranking introduced by Moretti et al. in [35]. We show that any social ranking can be encoded as an MC ranking, but that is not true the other way around. Therefore, the MC ranking is more general. Furthermore, we also show that our MC lex-cel generalises the lex-cel social ranking solution introduced by Bernardi et al. in [12].

The social ranking [35] considers a set of elements $X$, and a ranking over coalitions of these elements, namely a ranking over $\mathcal{P}(X)$. The purpose of a social ranking is to transform or ground this power set ranking into a ranking over $X$. Formally:

**Definition 11.** A social ranking is a function $sr : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$ which transforms a ranking over $\mathcal{P}(X)$ into a ranking over the elements of $X$.

The goal of a social ranking and of an MC ranking is the same: to obtain a ranking over $X$. Nonetheless, the starting points for the computation of the two rankings are different. While a social ranking considers a ranking over the power set of $X$, an MC ranking considers criteria, a ranking over criteria and a labelling relating elements to criteria. Note though that it is possible to define a function that transforms a social ranking into an MC ranking. Since the input of $sr$ is in $\mathcal{R}(\mathcal{P}(X))$ and the input of $mcr$ is in $\mathcal{R}(C) \times \mathcal{L}(X,C)$, we propose a function $t : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(C) \times \mathcal{L}(X,C)$ to transform the input of a social ranking into an input for MC ranking. Therefore, this transformation function is such that $t(\succeq_P) = (\lambda, \succeq_C)$. Let $X$ be a set of elements and $\succeq_P$ a ranking over $\mathcal{P}(X)$, we build function $t$ as follows:

1. We transform the sets in $\mathcal{P}(X)$ into criteria: $C = \{c_S, \forall S \in \mathcal{P}(X)\}$.
2. We obtain the ranking over criteria as a direct translation of the ranking over sets: $c_S \succeq_C c_{S'} \Leftrightarrow S \succeq_P S'$.
3. Finally, to define a labelling function, note that a social ranking does not consider gradings. However, we can consider one label to indicate that an element aligns with criterion $c_S$ (the element appears in set $S$), and another label to indicate that the element is neutral with regard to this criterion (the element does not appear in $S$). To do so, we define labels $l_1$ and $l_0$ respectively (along with the unused $l_{-1}$ to make the label system symmetric). Hence, we define the label system $LS = \langle L, \geq_L \rangle$, with $L = \{l_1, l_0, l_{-1}\}$, and order $l_1 \geq_L l_0 \geq_L l_{-1}$. Then, we build a labelling $\lambda$ that specifies whether an element $x$ is related to $c_S$ with label $l_1$ if $x \in S$, or with label $l_0$ if $x \notin S$:

$$\lambda(x, c_S) = \begin{cases} l_1, & \text{if } x \in S \\ l_0, & \text{if } x \notin S \end{cases} \tag{8}$$

The $t$ function allows us to transform any social ranking input into an MC ranking input. In fact, in what follows we prove that MC rankings generalise social rankings. Before that, we need an auxiliary result regarding the properties of function $t$ as shown by the following lemma.

**Lemma 2.** *The* $t$ *function is injective, but not exhaustive.*

**Proof.** Suppose that $t$ is not injective. Thus, for a given power set $\mathcal{P}(X)$, there are two different rankings $\succeq, \succeq' \in \mathcal{R}(\mathcal{P}(X))$, such that $t(\succeq) = t(\succeq')$. Since $\succeq, \succeq'$ are different rankings, $\exists Y, Z \in \mathcal{P}(X)$, such that $Z \succeq Y$, while $Z \not\succeq' Y$. Note though that in these cases when applying $t$, we would have that $c_Z \succeq_C c_Y$ and $c_Z \not\succeq'_C c_Y$, which contradicts the assumption that $t(\succeq) = t(\succeq')$. Thus, $t$ is injective. In terms of exhaustivity, $t$ is not exhaustive because labellings using labels other than $l_1$ and $l_0$ can never be the image of a social ranking.

Thanks to Lemma 2, we prove our first general result.

**Theorem 2.** *MC rankings generalise social rankings. That is, given a set of elements $X$, a power set $\mathcal{P}(X)$, a ranking over the power set $\succeq_P$, and a social ranking $sr : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$, there exists an MC ranking $mcr$, such that $sr(\succeq_P) = mcr(t(\succeq_P))$, but the reverse does not hold in general.*

**Proof.** To prove the theorem we have to find a $mcr$ function such that $sr(\succeq_P) = mcr(t(\succeq_P))$. Consider $mcr = sr \circ t^{-1}$. In this case, we would have that $mcr(t(\succeq_P)) = sr(t^{-1}(t(\succeq_P))) = sr(\succeq_P)$. In the previous lemma we have seen that $t$ is injective but not exhaustive in general, meaning that in general it is not invertible. Note though that $t$ is invertible when restricted to the domain $t(\mathcal{R}(X))$. In this case, since we start in this domain, $t^{-1}$ exists, meaning that $mcr = sr \circ t^{-1}$ is a valid function which proves the theorem.

This last theorem proves that all social rankings can be cast as an equivalent MC ranking. Also, since $t$ is not exhaustive there are many MC rankings that cannot be cast as social rankings, meaning that the MC ranking is more general. Regarding this last result, an interesting question we have to address is the relation between MC lex-cel and lex-cel (see Section 2). The next theorem shows that MC lex-cel generalises lex-cel.

**Theorem 3.** *MC lex-cel generalises lex-cel, that is, given a set $X$ and a ranking $\succeq_P$ over $\mathcal{P}(X)$, $mclex(t(\succeq_P)) = lex(\succeq_P)$.*

**Proof.** Suppose that $lex(\succeq_P) = \succeq_e$ and $mclex(t(\succeq_P)) = \succeq'_e$. We will see that given $x, y \in X$, $x \succeq_e y \Leftrightarrow x \succeq'_e y$. We start with $x \succeq_e y \Rightarrow x \succeq'_e y$. First, suppose that $x \succ_e y$ ($x \succeq_e y$ and $x \nsim_e y$). Then, from the definition of lex-cel, we would have that $\theta(x) >_{lex} \theta(y)$. Now, suppose that $\mathcal{P}(X)/\sim_P = \{\Sigma_1, \dots \Sigma_k\}$ with quotient order $\Sigma_1 \succ_P \dots \Sigma_k$. Then, $\theta(x) = (|S \in \Sigma_1 : x \in S|, \dots, |S \in \Sigma_k : x \in S|)$ and $\theta(y) = (|S \in \Sigma_1 : y \in S|, \dots, |S \in \Sigma_k : y \in S|)$. Hence, $\theta(x) >_{lex} \theta(y)$ means that $\exists\ \Sigma_i$ such that $|S \in \Sigma_i : x \in S| > |S \in \Sigma_i : y \in S|$, and $\forall\ \Sigma_j \succ_P \Sigma_i$, $|S \in \Sigma_i : x \in S| = |S \in \Sigma_i : y \in S|$. By applying $t$ to $\succeq_P$, we obtain that $\succeq_C$, such that any sets $S, S' \in \mathcal{P}(X)$ are transformed into criteria $c_S, c_{S'} \in C$, and $\succeq_P$ is transformed into $\succeq_C$ following $S \succeq_P S' \Leftrightarrow c_S \succeq_C c_{S'}$. Hence, the image for $t$ of each equivalence class $\Sigma_i \in \mathcal{P}(X)/\sim_P$ is a criterion equivalence class $\kappa_i \in C/\sim_C$, and the quotient order then satisfies that $\Sigma_i \succ_P \Sigma_j \Leftrightarrow \kappa_i \succ_C \kappa_j$. Recall that the labelling obtained by $t$ is built following $\lambda(x, c_S) = l_1 \Leftrightarrow x \in S$. Thus, $|S \in \Sigma_i : x \in S| > |S \in \Sigma_i : y \in S|$ implies that $|c \in \kappa_i : \lambda(x,c) = l_1| > |c \in \kappa_i : \lambda(y,c) = l_1|$. Similarly, $\forall\ \Sigma_j \succ_P \Sigma_i$, $|S \in \Sigma_i : x \in S| = |S \in \Sigma_i : y \in S|$ implies that $\forall \kappa_j \succ_C \kappa_i$, $|c \in \kappa_j : \lambda(x,c) = l_1| = |c \in \kappa_j : \lambda(y,c) = l_1|$. Note that, in this case, since the label system only contains $l_1, l_0$, and $l_{-1}$, we have that $s_{max} = 1$, and hence $\forall \kappa \in C/\sim_C$, $\mu(x,\kappa) = (na(x,\kappa,1))$. Moreover, from the definition of $t$, the labelling does not assign $l_{-1}$. Therefore, we have that $na(x,\kappa,1) = |c \in \kappa : \lambda(x,c) = l_1|$, and overall $\mu(x,\kappa) = (|c \in \kappa : \lambda(x,c) = l_1|)$. Now, we have that $|c \in \kappa_i : \lambda(x,c) = l_1| > |c \in \kappa_i : \lambda(y,c) = l_1|$, implying that $\mu(x,\kappa_i) >_{lex} \mu(y,\kappa_i)$, and $\forall \kappa_j \succ_C \kappa_i$, $|c \in \kappa_j : \lambda(x,c) = l_1| = |c \in \kappa_j : \lambda(y,c) = l_1|$, which implies that $\mu(x,\kappa_j) = \mu(y,\kappa_j)$. This is precisely the definition of $\mu(x) \succ_{lex} \mu(y)$, which means that $x \succ'_e y$. Similarly, if $x \sim_e y$, $\theta(x) = \theta(y)$. Therefore, $\forall i$, $|S \in \Sigma_i : x \in S| = |S \in \Sigma_i : x \in S|$. If we apply $t$, this means that $\forall i$, $|c \in \kappa_i : \lambda(x,c) = l_1| = |c \in \kappa_i : \lambda(y,c) = l_1|$, and then $\mu(x) = \mu(y)$, hence following that $x \sim'_e y$.

When it comes to the reverse implication, $x \succeq'_e y \Rightarrow x \succeq_e y$, suppose that $x \succ'_e y$. In this case, $x \preceq_e y$ cannot happen because we have seen above that it would imply that $x \preceq'_e y$, which is not true. Then, the only possibility is that $x \succ_e y$. We can follow the same reasoning to prove that $x \prec'_e y \Rightarrow x \prec_e y$ and $x \sim'_e y \Rightarrow x \sim_e y$.

## 7. Case study: a value alignment problem

The purpose of this section is to illustrate how MC-lexcel can be used to solve a value alignment problem, that is, a decision-making problem where elements have to be chosen by considering their alignment with multiple moral values. In particular, we focus on the problem tackled by a decision maker (e.g. a policy maker) when tasked with selecting the collection of regulatory norms that are most closely aligned with the moral values of a society. In Section 7.1, we introduce the decision-making problem. In Section 7.2 we discuss how to exploit MC-lexcel to solve the decision-making problem computationally. Finally, in Section 7.3 we discuss a case study in a healthcare context, concerned with selecting norms related to hospital admission. Furthermore, we compare the qualitative solving method detailed in Section 7.2 with existing methods in the literature.

### 7.1. Defining the value alignment problem

Within societies, norms have long been used as a coordination mechanism [6]. On the one hand, the literature on Normative multi-agent systems has traditionally focused on establishing norms to regulate agents' behaviour by means of: emergence [40,47], an empirical bottom-up approach; off-line norm synthesis [1], a formal top-down approach; and on-line norm synthesis [32,33], which is empirical and top-down. We refer to the set of norms enacted in a society as a *norm system*. On the other hand, norms have also been related to *moral values*[5] [17], which are used as guiding criteria for the selection [46,45] or synthesis [31] of the norm system to be enacted. Indeed, composing a set of norms that promote ethical behaviour (i.e., moral values) naturally induces this ethical behaviour in the society. Moreover, if different moral values can be promoted, then it seems reasonable to prioritise the most preferred ones. Consider, for example, a government that enacts norms limiting pollution. In this case, we can confidently infer that this government prioritises sustainability over other values such as development.

However, the problem of selecting the regulatory norms that align best with the ethical principles of a society (or, in other words, the *most value-aligned norm system*) is not straightforward. In addition to the different values and preferences over them

[5] According to the Ethics literature, moral values express the moral objectives worth striving for [48].

that a society may have, we must also consider whether norms actually promote or demote those values, as well as the degree of promotion/demotion. Some of the literature in Philosophy discusses a number of these aspects [18]. Nonetheless, in the Artificial Intelligence literature, while value promotion and demotion are commonly considered, the degrees of such relations are not typically considered (e.g. [28], [11], [43]). In fact, to the best of our knowledge such aspects have only been considered in legal cases [10].

Against this background, we introduce our value alignment problem while considering promotion and demotion relationships between norms and values as first-class citizens. Thus, we first introduce the formal objects required for the problem, namely: norms, value system, and the relationships between norms and values.

We define the core notion of our problem, the norm, as a simplification of the one proposed by Lopez et al. in [50]. We start by considering a MAS (multi-agent system) with a set of agents $Ag$ that can perform actions in a finite set $\mathcal{A}$. Furthermore, we consider a propositional language $\mathcal{PL}$ (with propositions in $\mathcal{P}$ and the logical operator "and"), a set of states $S$, and a state transition function that changes the state of the world when agents perform actions (following the multi-agent system model introduced by Morales et al. in [32,34]). Then, a norm is composed of a precondition $\varphi \subseteq \mathcal{P}$ (with an "and" semantic between propositions), an action in $\mathcal{A}$, and a deontic operator $\theta$ to establish Obligations ($Obl$), Permissions ($Per$), and Prohibitions ($Prh$). With these definitions in place, we define a norm as:

**Definition 12** *(Norm).* A norm is a pair $\langle \varphi, \theta(a) \rangle$, where $\varphi$ is a precondition in the language $\mathcal{PL}$; $a \in \mathcal{A}$ is the regulated action, and $\theta \in \{Obl, Per, Prh\}$ is a deontic operator.

**Example 7.** Within a healthcare context, we may have a norm permitting the hospital admission of incoming patients: ⟨patient_in, Per(admit)⟩.

Let $N$ be a set of candidate norms; the norms in $N$ might have relationships between them [46]. We consider two types of such norm relations, namely norm exclusivity and norm generalisation and note them as $R_x$, and $R_g$ respectively (we assume the decision-maker has sufficient knowledge of the domain to detect and provide these norm relations). On the one hand, we say $n, n'$ are *exclusive norms*, noted as $(n, n') \in R_x$, when we cannot enact both of them simultaneously. On the other hand, we say they have a direct generalisation relation, noted $(n, n') \in R_g$, meaning $n$ is more general than $n'$. With regards to generalisation relations, we note as $S(n)$ and $A(n)$, the successors and ancestors of $n$ respectively. Formally:

**Definition 13.** Given a norm $n \in N$, its ancestors are the norms that (directly or indirectly) generalise it: $A(n) = \{n' \in N : \exists n_1, \dots, n_k,$ and $(n', n_1), \dots, (n_k, n) \in R_g\}$. Conversely, successors are the norms that are (directly or indirectly) generalised by $n$: $S(n) = \{n' \in N : \exists n_1, \dots, n_k,$ and $(n, n_1), \dots, (n_k, n') \in R_g\}$.

Norms and their relations form a structure called a norm net.

**Definition 14.** Let $N$ be a set of norms and $R = \{R_x, R_g\}$ the set of norm relations (exclusivity and generalisation), we call *norm net* the tuple $\langle N, R \rangle$.

**Definition 15.** We apply the term *norm system* to any subset $\Omega \subseteq N$.

Not all norm systems are of our interest; note that norm systems may have conflicts (if they contain exclusive norms) or redundancy (if they contain norms related through generalisation). Thus, we focus on *sound* norm systems, i.e. those that are conflict-free and non-redundant [46].

**Definition 16.** Let $\langle N, R \rangle$ be a norm net, then we consider a norm system $\Omega \subseteq N$ to be sound iff it is:

- Conflict-free: $\forall n_i, n_j \in \Omega$, $(n_i, n_j) \notin R_x$
- Non-redundant: $\forall n$, with $|\bar{S}(n)| > 1$, then $\bar{S}(n) \nsubseteq \Omega$.

Where $\bar{S}(n) = \{n' \in N, (n, n') \in R_g\}$ stands for the set of direct successors.

Moral values are principles deemed valuable by a society [48]. Ethical choice typically implies a set of moral values [13] and preferences regarding them. Indeed, some values are preferred over others [11], and these preferences must impact the decision-making process. For that reason, we consider the *value system* to be a structure formed by moral values and their preferences [11,28,46]. Thus, we say that the value system guides ethical reasoning. While several types of preferences have been used to formalise this structure, we favour rankings because they are the least restrictive preference structures satisfying totality. In this manner, given any pair of values, we can assert a preference between them (which may not be possible with non-total preferences such as partial orders). Therefore, we define the value system as follows.

**Definition 17.** Let $V$ be a non-empty set of moral values, and $\succeq_v$ a ranking over $V$, we call *value system* the tuple $\langle V, \succeq_v \rangle$.

Now we are ready to formalise how norms and values relate. To do so, we can leverage the notion of the label system $\langle L, >_l, \lambda \rangle$, introduced by Definition 3, with each label corresponding to either a certain degree of promotion or demotion, and with function $\lambda : N \times V \rightarrow L$ assigning a label to each norm-value pair. We impose a neutral label $l_0$ in $L$ to indicate that a norm and a value are unrelated. This label also sets the boundary between promoting and demoting labels: labels more preferred than $l_0$ are promoting labels, while those less preferred than $l_0$ are demoting labels. Notice that the sign function in equation (2) signals whether a label represents promotion (1), demotion (-1), or if it is neutral (0). Moreover, the strength function in equation (3) characterises the degree of promotion/demotion of labels. Thus, given a label $l \in L$, the more labels between $l$ and $l_0$, the larger its promotion/demotion degree (i.e., the stronger $l$ is).

Thanks to the objects formally introduced so far, we are ready to introduce our decision-making problem, the so-called *generalised value-aligned norm selection problem* (GVANS).[6] The input of the GVANS problem is: (i) a norm net $\langle N, R \rangle$; (ii) a value system $\langle V, \succeq_v \rangle$; and (iii) a symmetric label system $\langle L, >_l, \lambda \rangle$ that sets the relation between norms and values. Solving a GVANS problem consists in composing the sound norm system which best aligns with the value system, taking into account the degree of promotion/demotion of norm-value relations as expressed by the label system.

### 7.2. Solving the value alignment problem

When deciding on the most value-aligned norm system, we follow the following proposition: the more preferred the values promoted by a norm system, the more preferred the norm system, or, in other words, the more *value-aligned*. To obtain the most value-aligned norm system (i.e., to solve a GVANS problem) we will proceed in two steps.

First, we exploit MC-lexcel to obtain a ranking over individual norms from a ranking over values in a value system. This is straightforward if we consider that the values in $V$ act like criteria (i.e. $C = V$), and value preferences are cast over the elements of the decision (i.e. the norms in $N$). Importantly, our aim is to use the norm ranking to later select the set of norms that best aligns with the value system. Since norms can both promote and demote values, there might be norms which, overall, demote more preferred values than those they promote. We call these norms *non-beneficial* norms. In contrast, *beneficial* norms are those that promote more preferred values than those they demote. A simple informal way to differentiate between beneficial and non-beneficial norms is to compare them to a neutral norm $n_0$. We define $n_0$ as an artificial norm that is neutral with regards to all the moral values in the value system. Thus, informally:

**Definition 18.** A beneficial norm is a norm that is more preferred than $n_0$. Norms less preferred than or indifferently preferred to $n_0$ are non-beneficial norms. We note as $N_{ben} \subseteq N$ the subset of beneficial norms in $N$.

When selecting a set of norms, we want to select only beneficial norms and avoid non-beneficial norms. In other words, the solution to the GVANS problem is a set of norms in $N_{ben}$. With MC-lexcel we can obtain a ranking that allows us to compare norms, but we must also know which norms are beneficial and which are not. In line with Definition 18 above, we exploit MC profiles to differentiate between them. Thus, in the case of MC profiles:

**Definition 19.** We say that a norm $n \in N$ is beneficial if $\mu(n) >_{lex} \mu(n_0)$. On the other hand, a norm is non-beneficial if $\mu(n_0) \geq_{lex} \mu(n)$. Thus, in this case, $N_{ben} = \{n \in N : \mu(n) >_{lex} \mu(n_0)\}$.

Indeed, since we build the ranking from the MC profiles of norms, a norm that is less preferred than $n_0$ will be a norm whose MC-profile is worse than that of a totally neutral norm. This is the case when an MC profile contains more demotion labels than promotion labels, or contains demotion labels associated to more preferred values. Thus, by applying MC-lexcel considering $N \cup \{n_0\}$, we obtain a ranking $mclex(\succeq_v) = \succeq_n$ in which not only we can compare norms, but also $n_0$ partitions norms between beneficial (when $n \succ_n n_0$) and non-beneficial (when $n_0 \succeq_n n$) norms.

The next step is to use the norm ranking to compose the desired set of value-aligned norms. Since only beneficial norms should be taken into account when composing the norm set, we discard the non-beneficial norms hereafter. Hence, we now consider the ranking only over beneficial norms $\succeq_n^{ben}$ obtained from the MC ranking over all norms. We formalise this using the following restriction:

**Definition 20.** The restriction function *ben* is a function $ben : \mathcal{R}(N) \rightarrow \mathcal{R}(N_{ben})$, such that $\forall \succeq_n \in \mathcal{R}(N)$ and $\forall n_1, n_2 \in N_{ben}, ben(\succeq_n) = \succeq_n^{ben}$ is such that $\succeq_n^{ben} = \{(n_1, n_2) \in \succeq_n : n_1, n_2 \in N_{ben}\}$.

Our final step consists in transforming the ranking over beneficial norms into a ranking over norm systems. For that, we utilise the anti-lexcel operator introduced in [43]. Let $N_{ben}$ be a set of beneficial norms, and $\succeq_n^{ben}$ a ranking over these norms, the anti-lex-cel function $ale : \mathcal{R}(N_{ben}) \rightarrow \mathcal{R}(\mathcal{P}(N_{ben}))$ is a lifting function which generates a ranking over subsets of beneficial norms, namely over the norm systems in $\mathcal{P}(N_{ben})$. Therefore, the composition of MC-lexcel, the restriction to beneficial norms, and anti-lex-cel, transforms preferences over values into a value system for preferences over beneficial norm systems. We formally define this composition as follows:

[6] The GVANS problem is a generalisation of the VANS problem introduced in [43], which disregarded demotions and promotion relations with different degrees.

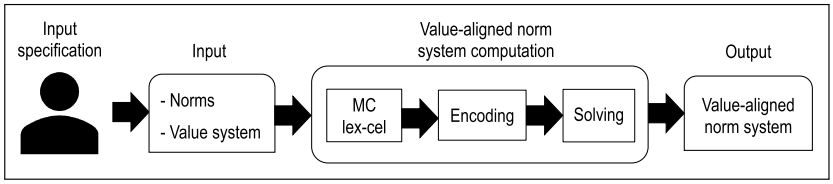


**Fig. 1.** Steps for the computational approach to maximise the value-alignment of norm systems.

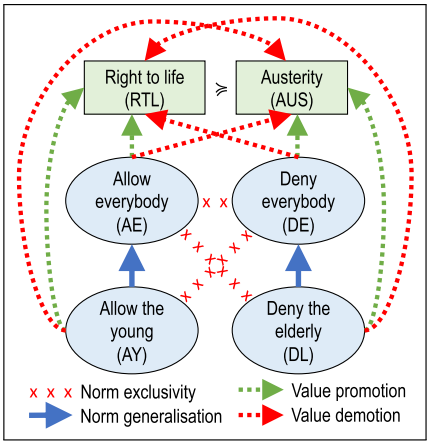


**Fig. 2.** Representation of the relationships between norms and norms and values in our healthcare case study.

**Definition 21.** We call $nsr : \mathcal{R}(V) \rightarrow \mathcal{R}(\mathcal{P}(N_{ben}))$ ($nsr$ for norm system ranking) the function $nsr = ale \circ ben \circ mclex$. Thus, for a value ranking $\succeq_v \in \mathcal{R}(V)$, $nsr(\succeq_v) = ale(ben(mclex(\succeq_v))) = \succeq$ is a ranking over norm systems (introduced in Definition 15) composed of beneficial norms.

The solution to the GVANS problem at hand will be the most preferred sound norm system produced by a norm system ranking. Unfortunately, although a norm systems ranking helps us obtain the solution to a GVANS problem, the cost of building a whole ranking over norm systems (elements in $\mathcal{P}(N_{ben})$) turns out to be rather costly. As discussed in [43], building the ranking using anti-lex-cel takes $O(2^{2^{|N|}})$ in the worst case (and when all norms are beneficial). Nonetheless, in [43] we show that it is possible to avoid the explicit computation of a whole ranking over norm systems. Indeed, in [43] we show how to cast the problem of selecting the most preferred norm system as an optimisation problem that can be encoded as a BIP (Binary Integer Program). This BIP only employs $|N|$ decision variables and can be solved with the aid of standard BIP solvers (e.g. CPLEX[7] or Gurobi[8]). We propose following the same approach here. Fig. 1 shows the steps of our computational approach to compute the most value-aligned norm system. First, given a set of norms and a value system as input, we apply MC-lexcel to obtain a norm ranking. After that, we restrict the norm ranking to beneficial norms, and then we use this beneficial norm ranking to do the encoding of the GVANS problem as a BIP. Finally, we solve the BIP with the aid of standard BIP solvers. Appendix A provides more details on this approach. We refer the reader to that appendix for details on encoding a GVANS problem.

Henceforth, we refer to the method outlined in Fig. 1 as the **qualitative approach with graded value promotion and demotion**. The next section illustrates and compares it to previous approaches.

### *7.3. Comparing solving methods*

Following Example 7 on healthcare, here we introduce a simple example that illustrates the **qualitative approach with graded value promotion and demotion** described in Section 7.2. Furthermore, we use it to compare our approach to those in the literature. On the one hand, [46] proposes a numerical approach that first assigns a utility to each norm – which represents their value alignment – and then selects norms by maximising their cumulative utility. Here, we show that asserting norm utilities may introduce biases that our qualitative approach avoids. On the other hand, although the work in [43] is also qualitative, it has limited expressiveness, since it does not allow for demotion, nor for different degrees of promotion/demotion. Overall, we show that, for specific cases, these other methods in the literature fail to produce a norm system that is the most closely aligned with the given moral values.

[7] IBM CPLEX optimiser: https://www.ibm.com/de-de/analytics/cplex-optimizer, last accessed Feb 2023.
[8] Gurobi: https://www.gurobi.com/, last accessed Feb 2023.

**Table 1**
Value-norm relationships for the three methods: $\lambda$ (our method); U (utilitarian [46]); and B (binary [43]).

| | AE | | | DE | | | AY | | | DL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\lambda$ | U | B | $\lambda$ | U | B | $\lambda$ | U | B | $\lambda$ | U | B |
| RTL | HP | 1 | 1 | HD | -1 | 0 | P | 0.7 | 1 | HD | -0.8 | 0 |
| AUS | HD | -1 | 0 | HP | 1 | 1 | D | -0.2 | 0 | HP | 0.8 | 1 |

As previously mentioned, our case study focuses on selecting norms related to hospital admission. In particular, as Fig. 2 shows, we consider four norms:

- AE = ⟨patient_in, Per(admit)⟩: Allow admission to Everybody;
- DE = ⟨patient_in, Prh(admit)⟩: Deny admission to Everybody;
- AY = ⟨young_patient_in, Per(admit)⟩: Allow admission to the Young;
- DL = ⟨elder_patient_in, Prh(admit)⟩: Deny admission to the eLderly.

Thus, $N = \{\text{AE, DE, AY, DL}\}$. Furthermore, as shown in Fig. 2, since admission cannot be allowed and denied simultaneously, some of these norms are exclusive: (AE, DE)$\in R_x$, (AY, DE)$\in R_x$, (AE, DL)$\in R_x$). Moreover, regulating the admission to everybody includes the young and the elders, and hence, AE generalises AY and DE generalises DL.

As for values, in this setting, we consider two moral values $V = \{\text{RTL, AUS}\}$: "Right To Life/medical care" (RTL) and "Austerity" (AUS), and a preference of RTL $\succeq_v$ AUS. Fig. 2 depicts how norms allowing admission promote RTL and demote AUS, whereas the norms denying it behave conversely. However, to express promotion/demotion degrees we consider a label system $\langle L, >_l, \lambda\rangle$ with the following labels: high promotion ($HP$), promotion ($P$), neutral ($l_0$), demotion ($D$) and high demotion ($HD$) ($L = \{HP, P, l_0, D, HD\}$) and linear order $HP >_l P >_l l_0 >_l D >_l HD$. Note that, e.g., $stg(HP) = 2$ and $sgn(HP) = 1$, whereas $stg(HD) = 2$, and $sgn(HD) = -1$. The $\lambda$ columns in Table 1 detail the $\lambda$ function that completes our label system. Overall, general norms that apply to everybody are strongly related to the values. This is also the case for elders, since they are most likely to require admission. Alternatively, we consider the relationship with young people to be less strong, since they are less likely to require admission.

From here, we apply our *qualitative approach with graded value promotion and demotion* to compute the norm ranking as AE $\succeq_n$ AY $\succeq_n n_0 \succeq_n$ DL $\succeq_n$ DE. This is because using Equation (5), we have $\mu(AE) = ((1,0),(-1,0))$, $\mu(AY) = ((0,1),(0,-1))$, $\mu(DL) = ((-1,0),(1,0))$, and $\mu(DE) = ((-1,0),(1,0))$. Note that DE and DL are non-beneficial norms because they are less preferred than $n_0$ (due to their demotion of the most preferred value). Therefore, by restricting the ranking to beneficial norms, we have that AE $\succeq_n^{ben}$ AY. Next, we obtain the norm system ranking $\{\text{AE, AY}\} \succeq \{\text{AE}\} \succeq \{\text{AY}\}$. However, {AE, AY} is not sound (see Definition 16) because AE generalises AY and, hence, the method will choose {AE} as the most value aligned norm system to be enacted. Indeed, considering that AE is the most general norm with the highest promotion of RTL – the most preferred value – then, {AE} stands for the expected solution.

Alternatively, when considering the **quantitative utilitarian approach** in [46], the task of assigning (and justifying) numerical degrees of promotion/demotion turns out to be more difficult. U columns in Table 1 detail the grades we use in this comparison. Extreme grades now become 1 and -1 respectively. AY promotes RTL with 0.7 and demotes AUS with -0.2 because the young are only a small portion of the admitted patients and the cost of their medical care is relatively low. DL demotes RTL with -0.8 and promotes AUS with 0.8 since most people at risk of dying are elders and they usually require the most expensive medical care.

Subsequently, the procedure in [46] computes norm utilities – for simplicity, we take the random $\epsilon \in (0,1]$ to be 0.5 – as: $u(\text{AE}) = 0.5$, $u(\text{DE}) = -1.5$, $u(\text{AY}) = 0.6$, $u(\text{DL}) = -1,2$ so that the sound norm system with the highest utility is {AY}. This means that the quantitative utilitarian method selects a norm system that fails to regulate admissions of elder people. This is because AE strongly demotes the AUS value, and this diminishes its utility.

If we now consider the **binary qualitative approach** used by [43], the B columns in Table 1 are limited to representing promotion (1) and no-promotion (0). This method produces a norm ranking of AE $\sim_n$ AY $\succeq_n$ DE $\sim_n$ DL. Notice that AE $\sim_n$AY because both norms promote RTL and the method is not expressive enough to capture different grades of promotion, even though admitting everybody (AE) is clearly a better norm (i.e., it is far more inclusive) than only admitting the young (AY). Consequently, this norm ranking leads to the following ranking of sound norm systems: {AY, DL} $\succeq$ {AE} $\sim$ {AY} $\succeq$ {DE} $\sim$ {DL}, where {AY, DL} supports both RTL and AUS values. Hence, the binary qualitative method selects the enactment of {AY, DL}, which fails to be aligned with the value system because denying admission to elders (DL) demotes the most preferred value, namely the right to life (RTL). The reason for considering such a undesirable norm is a direct consequence of its failure to capture demotion. In fact, it only selects norms based on their merits without considering their detrimental effects.

In conclusion, despite its simplicity, this example illustrates how our method overcomes the shortcomings of the aforementioned methods [46] and [43] in producing a norm system that is most closely aligned with the value system at hand. In fact, [43] already reported a flaw in [46] that can cause that a number of norms slightly promoting least preferred values to end up having more utility – and thus being chosen – than a single really useful norm if they are exclusive. Indeed, although most preferred values should prevail, the quantitative method also fails to capture the absolute preferences of the value system.

The advantages of our method are two-fold. First, its graded qualitative labels for promotion and demotion are much simpler to define – and less prone to biases – than numerical degrees, and it also provides far more expressiveness than binary promotion alone. Second, its ranking method captures the preferences of the value system in the selection of the norm system to enact.

**Table 2**
Example of 10 projects that were selected in the Barcelona Participatory Budgeting process with the number of votes they received, the strategic area they affect, and their cost (the projects are ordered by number of votes).

| | Project | Votes | Strategic area | Cost in euros |
|---|---|---|---|---|
| $p_1$ | Improve the Espanya Industrial park | 3.351 | Green spaces | 850.000 |
| $p_2$ | Reform Capmany sport facilities to include cricket | 2.890 | Sport | 1.600.000 |
| $p_3$ | Towards a greener Gràcia neighbourhood | 2.885 | Green spaces | 500.000 |
| $p_4$ | Recover the paths from Vallcarca to Collserola | 2.655 | Green spaces | 50.000 |
| $p_5$ | Pacify Sants antic | 2.483 | Pacification | 200.000 |
| $p_6$ | Turn La Farigola school yard into an open green area | 2.430 | Green spaces | 350.000 |
| $p_7$ | Give more value to the Roman wall | 1.530 | Monuments | 65.000 |
| $p_8$ | Adapt Àngels' square for children | 1.450 | Pacification | 150.000 |
| $p_9$ | Create an accessible IT service point | 1.325 | Culture | 50.000 |
| $p_{10}$ | Build a bike lane on Via Augusta | 1.009 | Mobility | 1.100.000 |

## 8. Discussion regarding applicability

Although the main contribution of this work is theoretical, the previous section presented a norm selection case study with the aim of illustrating its application. However, this should not preclude the reader from understanding that our method can be applied in other scenarios. Indeed, we already mentioned some of these potential scenarios in the introduction section (college admissions, committee selections, scholarship assignments, or diet design). Here we aim to go a step further and discuss in some detail how an alternative domain in the context of participatory democracy [37] may benefit from our approach.

As pointed out by the European Commission [41], democracy aims at the greater good of society, where good cannot only be measured in monetary terms, but also requires considering what citizens perceive to be valuable. As a consequence, the European Joint Research Centre has elaborated a policymaker's guide for the 21st century[9] stressing that policies need to take into account and reflect the values and concerns of citizens. Here we also follow this stance and focus on the scenario of participatory budgeting –a mechanism enacted by policymakers to implement projects proposed by citizens– and enrich it by explicitly using citizens' values as selection criteria.

Participatory budgeting constitutes a democratic approach to deciding the funding of public projects [7]. Recently, it has attracted a lot of attention because of its ability to encourage citizens' participation in politics and to promote open democracy. Participatory budgeting processes have been mostly, though not exclusively, adopted in cities across the world, including New York,[10] Toronto,[11] Buenos Aires,[12] Paris,[13] Madrid,[14] and Seoul.[15]

Despite empowering citizens, current implementations of participatory budgeting processes suffer from a major practical caveat related to their limited convening capacity [29]. As a consequence, their representativeness may be compromised if, for instance, educated younger citizens are more inclined to participate in online processes than other population segments, such as the elderly, who may be affected by the digital divide. This would lead to a bias in the selection process towards the preferences of the population segment that actively participates in the process. Our method can help to tackle this problem if project proposals are selected by considering as a criterion the common values shared by the population. This requires gathering citizens' values and then evaluating the alignment of the project proposals with those values. Although we consider the gathering of citizens' values to lie outside the scope of this paper, it is worth mentioning that the work of Liscio et al. [25,26] can be used to detect context-specific values, and preferences over these values can be gathered through surveys such as the European Values Study [15]and then aggregated by means of the state-of-the art method proposed by Lera-Leri et al. [24].

Here, we develop an alternative case study for our work by considering a participatory budgeting process in which citizens submit a set of project proposals to which we then apply our selection process based on citizens' values. We illustrate this by taking real data from the participatory budgeting process held by the Barcelona city council in 2021.[16] This participatory budget is a relevant example because the process had an overall budget of 30 million euros, attracted the participation of more than 64000 citizens, and resulted in the selection of 76 (out of 184) projects being funded. Table 2 lists the names, number of votes, strategic areas, and the costs of 10 of the projects that were finally selected. As we can observe, they are related to different strategic areas such as sports, mobility, and culture which, if prioritised, can be used as criteria.

Typically, the decision selection process that is applied is the rank and select method [44]. This method is based solely on the number of votes cast as it first ranks the projects in decreasing order of the number of votes received (as presented in Table 2) and then selects projects until the budget is spent. As an example, imagine Barcelona had only received these 10 project proposals and they had to decide which to select if the available budget was 2.5 million euros. In this case, we would select projects $p_1$ and $p_2$,

[9] https://publications.jrc.ec.europa.eu/repository/handle/JRC126150 (last visited on 5th Oct 2022).
[10] https://www.participatorybudgeting.org (last visited on 28th Sept 2022).
[11] https://www.torontohousing.ca/residents/getting-involved/participatory-budgeting (last visited on 28th Sept 2022).
[12] https://www.buenosaires.gob.ar/areas/hacienda/pp/introduccion.php?menu_id=6784 (last visited on 28th Sept 2022).
[13] https://budgetparticipatif.paris.fr/bp/jsp/site/Portal.jsp (last visited on 28th Sept 2022).
[14] https://decide.madrid.es/presupuestos (last visited on 28th Sept 2022).
[15] https://yesan.seoul.go.kr/intro/index.do (last visited on 28th Sept 2022).
[16] https://www.decidim.barcelona/processes/PressupostosParticipatius?locale=es (last visited on 28th Sept 2022).

with a cost of 2.45 million euros, and since the cost of the next project is greater than the remaining budget, we would stop selecting projects (note that this means that we would not even select project $p_4$, which could be funded with the remaining budget).

Unfortunately, participation in this process was around 4%, and we may therefore consider citizens' to be more appropriate for selecting which projects to fund. Thus, we can consider that strategic areas represent different values and we can imagine that the city council of Barcelona carries out opinion polls to learn about citizens' preferences with regard to these strategic areas. Now, say that, in this hypothetical case, we conclude that citizens prefer Green spaces above all, followed by Pacification, Culture, Mobility, Monuments, and, finally, Sports. In particular, this means that project 2 has received a lot of support thanks to the bias of the participant base, but this would not be the case if the preferences of the whole population were considered.

From this, we can apply MC-lex-cel to obtain a ranking of proposals considering the preferences expressed regarding strategic areas (our criteria here). In this case, since Table 2 links each project to a single area, the ranking of projects obtained through MC-lex-cel would rank highly those projects that are related to more preferred areas (and equally those that are within the same area), hence $p_1 \sim p_3 \sim p_4 \sim p_6 \succeq p_5 \sim p_8 \succeq p_9 \succeq p_{10} \succeq p_7 \succeq p_2$.

Appendix B provides the encoding for participatory budgeting, which in turn applies the norm selection encoding in Appendix A. The solution results in the selection of projects $p_1$, $p_3$, $p_4$, $p_5$, $p_6$, $p_7$, $p_8$, and $p_9$. Therefore, although $p_2$ has a large amount of votes, it is related to the least preferred area, and since it is the most expensive project, it is optimal to select the more closely aligned (and less expensive) projects.

Finally, note that Barcelona's Participatory Budget related each project to only one strategic area. Nonetheless, in reality, these projects will probably be related to more than one area to various degrees (e.g. $p_1$ is concerned with improving a park, therefore it is highly related to the area of Green Spaces, but it may also be related to Pacification to a lesser degree). If this information was available, MC lex-cel could consider it to build the ranking of proposals.

## 9. Conclusions

In this paper, we have tried to make headway in supporting decision-makers that are challenged with comparing, and ultimately ranking, elements with regard to how such elements satisfy multiple criteria and how such criteria are preferred. This calls for a new decision-making framework, which we have formally introduced here. Our framework is based on a novel method for ranking single elements.

Ranking functions have been widely used to transform rankings. For instance, the social ranking function transforms a ranking over sets of elements into a ranking over the elements themselves. This paper contributes to the state of the art by proposing a novel ranking – the multi-criteria (MC) ranking lrev – and the function that creates them – the MC lex-cel – to transform complex preference (criteria) information into a neat and clear ranking of individual elements. Furthermore, we have positioned our findings with respect to the current literature by showing that our MC ranking generalises the social ranking and MC lex-cel generalises the lex-cel social ranking function and embodies dominance.

Finally, the paper also illustrates how they can be employed to help a decision-maker to tackle an ethical decision-making problem. Specifically, we define the Generalised Value-Aligned Norm Selection (GVANS) problem and solve it with a qualitative approach with graded value promotion and demotion. Overall, this method overcomes the shortcomings of previous methods, resulting in a norm system that is most closely aligned with the value system at hand.

In future work, we plan to study in more detail the properties of MC rankings and provide an axiomatisation for MC lex-cel. Indeed, a full axiomatisation is not straightforward, since the use of labels requires the definition of new properties outside of classic social choice ones.

## CRediT authorship contribution statement

**Marc Serramia:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Maite Lopez-Sanchez:** Supervision, Writing – original draft, Writing – review & editing. **Stefano Moretti:** Validation, Writing – original draft, Writing – review & editing. **Juan A. Rodriguez-Aguilar:** Supervision, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

No data was used for the research described in the article.

## Acknowledgements

Research supported by projects: COREDEM (H2020-785907); Crowd4SDG (H2020-872944); CI-SUSTAIN (PID2019-104156GB-I00); TAILOR (H2020-952215); COMRDI18-1-0010-02; 2021 SGR 00754; 2021 SGR 000313; THEMIS (ANR-20-CE23-0018); "FairTransNLP-Language: Analysing toxicity and stereotypes in language for unbiased, fair and transparent systems" (PID2021-

124361OB-C33) funded by MCIN/AEI/10.13039/501100011033/FEDER,UE; "AUTODEMO: The stealth side of participatory democracy: process preferences towards automated decision-making" code SR21-00329; "VAE: Value-Awareness Engineering" TED2021-131295B-C31 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR; and VALAWAI project, funded by the European Commission under Grant #101070930.

## Appendix A. A BIP encoding to solve the GVANS problem

Consider a set of norms $N$; a value system with a set of values $V$ and value preferences $\succeq_v$; a (symmetric) label system with set of labels $L$ and linear order $\geq_L$; and a labelling function $\lambda$ relating norms to values with labels in $L$. As explained in [43], we solve the GVANS problem by transforming it into an optimisation problem. This allows us to have an optimal solution without the cost of applying anti-lex-cel. To do so, we first apply MC-lex-cel to obtain a ranking over single norms. Suppose $mclex(\succeq_v) = \succeq$ is the ranking over $N$ based on the relations between norms and values, their labels, and the value preferences. Using this ranking we can produce a BIP encoding. In particular, we adapt the work in [43] and compute the equivalence classes in $\Xi_1, \dots, \Xi_r \in N/\succeq$ (with $\Xi_1 \succ \cdots \succ \Xi_r$) to numerically compute the value alignment of any set of norms $S \subseteq N$ with the so-called preference function $\mathfrak{p}$. This function consists of the sum of the preference over the equivalence class each norm is in. Therefore, the more norms and the more preferred they are, the greater the value-alignment ($\mathfrak{p}$) of subset $S$. Formally, the value alignment of S is:

$$\mathfrak{p}(S) = \sum_{i=1}^{r} |S \cap \Xi_i| \left( \sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1 \right), \text{ where } \mathfrak{p}(\Xi_r) = |\Xi_r|. \tag{A.1}$$

To find the set of norms $S$ that maximises value-alignment, [43] proves that we can use a BIP encoding that uses $|N|$ binary decision variables, one per norm. Decision variable $d_i$ associated to norm $n_i \in N$ is used for deciding whether the norm is selected or not. The objective function of the BIP is a weighted combination of the decision variables of norms. Each decision variable $d_i$ is weighted by a factor that depends on the preference $\mathfrak{p}$ of the equivalence class of the norm. Then, the objective function to maximise is:

$$\sum_{i=1}^{r} \sum_{n_w \in \Xi_i} d_w \left( \sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1 \right) \tag{A.2}$$

To ensure that we select a sound norm system, the BIP encoding must also consider the following constraints:

- Mutually exclusive (incompatible) norms cannot be selected at once:

$$d_i + d_j \leq 1 \text{ for each } (n_i, n_j) \in R_x \tag{A.3}$$

- All successors of a norm ($\bar{S}(n) = \{n' \in N, (n, n') \in R_g\}$) cannot be jointly selected:

$$\sum_{i=1}^{|\bar{S}(n)|} d_i \leq |\bar{S}(n)| \text{ for each } n \in N. \tag{A.4}$$

## Appendix B. BIP encoding to solve the selection of participatory budget projects

To select participatory budget projects, we apply the BIP encoding in Appendix A to consider preferences regarding projects. Consider a set of projects $X$ and $\succeq$ preferences over the projects as a ranking, with $\Xi_1, \dots \Xi_r \in X/\succeq$ and $\Xi_1 \succ \cdots \succ \Xi_r$. We want to maximise the preferences of the selected projects. The BIP encoding uses decision variables $d_i \in \{0, 1\}$ (0 for not selected, 1 for selected) for each $x_i \in X$. The target function is the same formula as in Appendix A, which is the sum of decision variables $d_i$ each multiplied by a parameter related to the equivalence class $x_i$ is in (see equation (A.2)).

The BIP would maximise this target function and if there are exclusivity or generalisation relationships between projects, this BIP encoding will also include the corresponding constraints. Similarly, budget constraints may also be added.

**Example 8.** Following the participatory budgets example in Section 8, we recall having preferences

$$p_1 \sim p_3 \sim p_4 \sim p_6 \succeq p_5 \sim p_8 \succeq p_9 \succeq p_{10} \succeq p_7 \succeq p_2$$

In this case, we have six equivalence classes $\Xi_1 \succ \cdots \succ \Xi_6$. To produce the encoding, first we aim to define the target function following Equation (A.2). In this case, applying the preference function defined in Equation (A.1), we have

$\mathfrak{p}(\Xi_6) = 1$, $\mathfrak{p}(\Xi_5) = 2$, $\mathfrak{p}(\Xi_4) = 4$, $\mathfrak{p}(\Xi_3) = 8$, and $\mathfrak{p}(\Xi_2) = 32$.

Note that to define this function we need $\mathfrak{p}(\Xi_i)$ for $2 \geq i \geq 6$ so we do not need $\mathfrak{p}(\Xi_1)$. Therefore, the target function (as per Equation (A.2)) is encoded as follows:

$48d_1 + 1d_2 + 48d_3 + 48d_4 + 16d_5 + 48d_6 + 2d_7 + 16d_8 + 8d_9 + 4d_{10}$

In this case, we have to consider the constraint of the budget, thus the sum of the costs of the selected proposals cannot surpass 2.5 million euros, meaning that we have to consider the constraint:

$850000d_1 + 1600000d_2 + 500000d_3 + 50000d_4 + 200000d_5 + 350000d_6 + 65000d_7 + 150000d_8 + 50000d_9 + 1100000d_{10} <= 2500000$

We solve this problem by using CPLEX and obtain the solution $\{p_1, p_3, p_4, p_5, p_6, p_7, p_8, p_9\}$.

## References


[1] Thomas Ågotnes, Wiebe Van Der Hoek, Carles Sierra, Michael Wooldridge, On the logic of normative systems, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI '07, 2007, pp. 1175–1180.
[2] Juan A. Aledo, José A. Gámez, Alejandro Rosete, A highly scalable algorithm for weak rankings aggregation, Inf. Sci. 570 (2021) 144–171.
[3] Encarnación Algaba, Stefano Moretti, Eric Rémila, Philippe Solal, Lexicographic solutions for coalitional rankings, Soc. Choice Welf. 57 (2021) 817–849.
[4] Tahar Allouche, Bruno Escoffier, Stefano Moretti, Meltem Ozturk, Social ranking manipulability for the cp-majority, Banzhaf and lexicographic excellence solutions, in: 29th International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}, 2020, pp. 17–23.
[5] Ritxar Arlegi, A note on Bossert, Pattanaik and Xu's "Choice under complete uncertainty: axiomatic characterization of some decision rules", Econ. Theory 22 (1) (2003) 219–225.
[6] Ofer H. Azar, What sustains social norms and how they evolve?: the case of tipping, J. Econ. Behav. Organ. 54 (1) (2004) 49–64.
[7] Haris Aziz, Nisarg Shah, Participatory budgeting: models and approaches, in: Tamás Rudas, Gábor Péli (Eds.), Pathways Between Social Science and Computational Social Science: Theories, Methods, and Interpretations, Springer International Publishing, Cham, 2021, pp. 215–236.
[8] Salvador Barberà, Walter Bossert, Prasanta K. Pattanaik, Ranking sets of objects, in: Handbook of Utility Theory, Springer, 2004, pp. 893–977.
[9] Sylvain Béal, Eric Rémila, Philippe Solal, Lexicographic solutions for coalitional rankings based on individual and collective performances, J. Math. Econ. 102 (2022) 102738.
[10] Trevor Bench-Capon, Henry Prakken, Adam Wyner, Katie Atkinson, Argument schemes for reasoning with legal cases using values, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, 2013, pp. 13–22.
[11] Trevor J.M. Bench-Capon, Katie Atkinson, Abstract argumentation and values, in: Argumentation in Artificial Intelligence, Springer, 2009, pp. 45–64.
[12] Giulia Bernardi, Roberto Lucchetti, Stefano Moretti, Ranking objects from a preference relation over their subsets, Soc. Choice Welf. 52 (4) (Apr 2019) 589–606.
[13] David Cooper, Value Pluralism and Ethical Choice, St. Martin Press, Inc., 1993.
[14] Jean-Paul Doignon, Stefano Moretti, Meltem Oztürk, On the ordinal invariance of power indices on coalitional games, in: 31st International Joint Conference on Artificial Intelligence, 2022.
[15] EVS, European values study, https://europeanvaluesstudy.eu, 2023, last visited on Febr. 2023.
[16] Peter C. Fishburn, Exceptional paper—lexicographic orders, utilities and decision rules: a survey, Manag. Sci. 20 (11) (1974) 1442–1471.
[17] Sven Ove Hansson, The Structure of Values and Norms, Cambridge University Press, 2001.
[18] Sven Ove Hansson, Formal Investigations of Value, Springer International Publishing, Cham, 2018, pp. 499–522, chapter 6.
[19] Adrian Haret, Hossein Khani, Stefano Moretti, Meltem Öztürk, Ceteris paribus majority for social ranking, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 303–309.
[20] Hossein Khani, Stefano Moretti, Meltem Öztürk, An ordinal Banzhaf index for social ranking, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, in: International Joint Conferences on Artificial Intelligence Organization, vol. 7, 2019, pp. 378–384.
[21] Christian Klamler, Ulrich Pferschy, Stefan Ruzika, Committee selection with a weight constraint based on lexicographic rankings of individuals, in: Francesca Rossi, Alexis Tsoukias (Eds.), Algorithmic Decision Theory, Springer, Berlin, Heidelberg, 2009, pp. 50–61.
[22] Sébastien Konieczny, Stefano Moretti, Ariane Ravier, Paolo Viappiani, Selecting the most relevant elements from a ranking over sets, in: Scalable Uncertainty Management: 15th International Conference, SUM 2022, Paris, France, October 17–19, 2022, Proceedings, Springer, 2022, pp. 172–185.
[23] Jérôme Lang, Jérôme Mengin, Lirong Xia, Voting on multi-issue domains with conditionally lexicographic preferences, Artif. Intell. 265 (2018) 18–44.
[24] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, Juan Rodriguez-Aguilar, Towards pluralistic value alignment: aggregating value systems through lp-regression, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, 2022, pp. 780–788.
[25] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, Pradeep K. Murukannaiah, Axies: identifying and evaluating context-specific values, in: Proceedings of International Conference on Autonomous Agents and MultiAgent Systems, 2021, pp. 799–808.
[26] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Pradeep K. Murukannaiah, What values should an agent align with? An empirical comparison of general and context-specific values, Auton. Agents Multi-Agent Syst. 36 (1) (Apr 2022).
[27] Roberto Lucchetti, Stefano Moretti, Tommaso Rea, Coalition formation games and social ranking solutions, in: 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Proceedings, New Zealand, May 2022 (Online).
[28] Jieting Luo, John-Jules Meyer, Max Knobbout, Reasoning about opportunistic propensity in multi-agent systems, in: AAMAS 2017 Workshops, Best Papers, 2017, pp. 1–16.
[29] Ank Michels, Laurens De Graaf, Examining citizen participation: local participatory policy making and democracy, Local Gov. Stud. 36 (4) (2010) 477–491.
[30] Grzegorz Miebs, Miłosz Kadziński, Heuristic algorithms for aggregation of incomplete rankings in multiple criteria group decision making, Inf. Sci. 560 (2021) 107–136.
[31] Nieves Montes, Carles Sierra, Synthesis and properties of optimally value-aligned normative systems, J. Artif. Intell. Res. 74 (sep 2022).
[32] Javier Morales, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Wamberto Vasconcelos, Michael Wooldridge, On-line automated synthesis of compact normative systems, ACM Trans. Auton. Adapt. Syst. 10 (1) (March 2015) 2:1–2:33.
[33] Javier Morales, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Michael Wooldridge, Wamberto Vasconcelos, Automated synthesis of normative systems, in: AAMAS 2013, 2013, pp. 483–490.
[34] Javier Morales, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Michael Wooldridge, Wamberto Vasconcelos, Synthesising liberal normative systems, in: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '15, 2015, pp. 433–441.
[35] Stefano Moretti, Meltem Öztürk, Some axiomatic and algorithmic perspectives on the social ranking problem, in: International Conference on Algorithmic Decision Theory, Springer, 2017, pp. 166–181.
[36] Tomoko Nemoto, David Beglar, Likert-scale questionnaires, in: JALT 2013 Conference Proceedings, 2014, pp. 1–8.
[37] William Nylen, L. Dodd, Participatory Democracy Versus Elitist Democracy, Springer, 2003.
[38] Prasanta K. Pattanaik, Bezalel Peleg, An axiomatic characterization of the lexicographic maximin extension of an ordering over a set to the power set, Soc. Choice Welf. 1 (2) (1984) 113–122.
[39] Alvin E. Roth, Marilda Sotomayor, Two-sided matching, in: Handbook of Game Theory with Economic Applications, vol. 1, 1992, pp. 485–541.
[40] Bastin Tony Roy Savarimuthu, Maryam Purvis, Stephen Cranefield, Martin Purvis, Mechanisms for norm emergence in multiagent societies, in: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '07, ACM, New York, NY, USA, 2007, pp. 173:1–173:3.
[41] Mario Scharfbillig, Laura Smillie, David Mair, Marta Sienkiewicz, Julian Keimer, Raquel Pinho dos Santos, Hélder Vinagreiro Alves, Elisa Vecchione, Laurenz Scheunemann, Values and Identities—A Policymaker's Guide, Publications Office of the European Union, 2021.
[42] Michael Schmitt, Laura Martignon, Dana Ron, On the complexity of learning lexicographic strategies, J. Mach. Learn. Res. 7 (1) (2006).
[43] Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, A qualitative approach to composing value-aligned norm systems, in: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, 2020, pp. 1233–1241.

[44] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Patricio Escobar, Optimising participatory budget allocation: the decidim use case, in: Artificial Intelligence Research and Development - Proceedings of the 22nd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2019, Mallorca, Spain, 23-25 October 2019, in: Frontiers in Artificial Intelligence and Applications, vol. 319, IOS Press, 2019, pp. 193–202.
[45] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Javier Morales, Michael Wooldridge, Carlos Ansotegui, Exploiting moral values to choose the right norms, in: Proceedings of the 1st Conference on Artificial Intelligence, Ethics and Society (AIES'18), 2018, pp. 1–7.
[46] Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Manel Rodriguez, Michael Wooldridge, Javier Morales, Carlos Ansotegui, Moral values in norm decision making, in: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18), International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 1294–1302.
[47] Toshiharu Sugawara, Emergence and stability of social conventions in conflict situations, in: Toby Walsh (Ed.), IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16–22, 2011, IJCAI/AAAI, 2011, pp. 371–378.
[48] Ibo van de Poel, Lambèr Royakkers, Ethics, Technology, and Engineering: An Introduction, Wiley-Blackwell, 2011.
[49] Steven Walczak, Terry Sincich, A comparative analysis of regression and neural networks for university admissions, Inf. Sci. 119 (1) (1999) 1–20.
[50] Fabiola López y López, Michael Luck, Mark d'Inverno, Constraining autonomy through norms, in: AAMAS, ACM, 2002, pp. 674–681.