Modelling Integration of Responsible AI Values for Ethical Decision-making

Abstract.

A principal requirement to achieve trustworthy-AI is to consider ethical aspects in the design and development of AI systems. This is particularly challenging when it comes to automatic decision-making, and requires appropriate tools to incorporate these aspects in the decision process. One way to address this issue is to evaluate the available alternatives in terms of their adherence to AI-specific moral values, i.e. responsible-AI values. In this article, we propose a hierarchical model to integrate responsible AI values as decision criteria in a ranking process. Each criterion represents a feature of a moral value that can evaluate alternatives on an ordinal scale, represented by a preference relation. We investigate different approaches for aggregating preferences and propose applicable methods based on an assumption of the priority among criteria. As a proof of concept, we adopt our framework on a problem of ranking several recommendation systems according to their adherence to responsible AI values. We use Answer Set Programming (ASP) to formally represent the implementation of our framework on this problem and derive a ranking of alternatives by logical reasoning.

1 Introduction

As the artificial intelligence advances, it is being increasingly used in applications that affect individuals and/or society. These applications range from medical diagnostics and criminal investigations to recruitment and recommendations. In such cases, we need to take ethical considerations into account in the decision process to enhance the liability of AI systems.

Current works in computational ethics can be divided mainly to deontological and/or utilitarianist approaches. The former views morality as duties and obligations that must be fulfilled based on a deontological principles. One of the main issues in this case is identifying norms and representing them formally. Utilitarianists view morality as the well-being or utility in the consequence of actions. Utility can be used as base to verify how good is an option comparing to others and identify the best one according to a utilitarianist principle e.g. act utilitarianism. A principal issue in this case which we are particularly interested in is to define a proper scale based on an explicit notion of the good that allows options to be compared reasonably.

One way to address this is to consider a pluralistic set of moral values in decision making process. Moral values and principles are a central topic in moral philosophy; in particular, value pluralism concerns questions about the existence of several distinct values or whether they are reducible to a super value[27]. Moral theories can be pluralistic according to their view of morality [21].

We refer to certain moral values that should be respected by AI autonomous agent as responsible-AI values. Many of the ethics guidelines discuss these values e.g. privacy, fairness, explainablity, nonmaleficence etc. [15],[1]. These values must be respected at all stages of AI design and development to ensure the its responsible use. In order to compare available options by considering responsible AI values in an automatic setting, we need to incorporate these values in the reasoning and decision process.

Several works on computational ethics have used a pluralistic view in their implementation of utilitarianism theories [3], [4]. These methods evaluate goodness by quantifying it on a cardinal scale and combine them by weightings and arithmetic calculations for decision making. They use elicited values for specific cases [19] or use random numbers to explore ethical dilemmas without proposing a method to actually obtain them [10], [6]. In certain cases, it is possible to compare options rationally by quantification. However, it may add arbitrary variables to the process that can bias the decision and lead to counter-intuitive results. Furthermore, there are other issues related to the representation and formalization of moral values that make their integration challenging; Moral values may represent abstract concepts or may have broad meanings [31] e.g. fairness. As a consequence, there is no agreement on a fixed list of AI values [25] or a universal interpretation of them. Last but not least, deriving the evaluative criteria from fundamental principles is not always feasible [22].

In this article we propose a framework based on multiple criteria with a hierarchical structure to integrate responsible AI values in decision making. The hierarchical criteria structure provides an explicit and explanatory representation of moral values. Each criterion represents a certain aspect of a moral value that can be used to rank alternatives on an ordinal scale, represented by a preference relation. An ordinal preference, allows logical reasoning that is essential for ethical decision-making. Furthermore, in many cases, an intuition of which option is better than another is much easier to ascertain than a precise quantification of the extent of this preference. We then investigate several approaches for aggregating preferences and propose applicable methods based on the assumption of the priority among criteria.

Lastly, we adopt our framework on a case study, where the goal is to ethically rank a set of recommender algorithms according to their adherence to responsible AI values. We show the process to design the hierarchical structure by making use of the AI ethics guidelines and literature. To implement this use case, we make use of Answer Set Programming(ASP) [17] which is a knowledge representation and reasoning paradigm with expressive formalism and efficient solvers. We use Clingo [12], as an answer set solver, for ASP.

The rest of the paper is organized as follows. Section 2 presents related works. In Section 3 we explain a formalization of our framework, and propose some aggregation approaches in Section 4. In Section 5 we describes a case study by adopting our framework for ethical ranking of several recommender systems. In Section 6 we discuss the conclusions and future perspectives.

2 Related Works

A line of research in computational ethics is modelling moral theories like consequnetialt or deontological theories in order to judge ethical permissibly of agent actions [28]. Consequentialist theories e.g. act utilitarianism often require a notion of utility to identify the best possible action. Most of the current works usually propose a quantified model for evaluation of utility which is either supposed to be given [19], elicited form a source[18], or is obtained by weighting and summing of utilities according to some components like moral values [6]. Quantification and arithmetic aggregation can be reasonable in some cases however it may cause counter-intuitive results in general. We avoid this issue in our framework by modelling the evaluation of alternatives as an ordinal preference relation. Ranking alternatives or modelling evaluations by preference relations is a common way of modelling in decision making problems. For example [20] use Conditional Preference networks (CP-Nets) for ethical decision making by comparing a distance between already evaluated alternatives(states). CP-nets provide a compact way to qualitatively model preferences over decisions with a combinatorial structure [7]. In an other work [14] propose a ranking of alternatives(sequence of actions or plans) by combining multiple ethical features assigned by moral principles with predefined importance. However none of these works allows an evaluation or ranking of alternatives based on moral values.

We have drawn inspirations from multi-criteria decision making(MCDM) methods for the design of our framework. These approaches are used to analyse and solve decision problems involving multiple criteria [13], [5]. All MCDM methods propose a way to combine the evaluation of criteria over alternatives. The alternatives can be evaluated in several ways, e.g. pairwise comparison of alternatives [11], approval voting [8], ordinal rankings of alternatives [16], and sorting classes of alternatives [24], [32]. Our model relies on a hierarchy of criteria as in Analytic Hierarchy Process (AHP)[29]. AHP is used to derive relative priorities from both discrete and continuous paired comparisons in multilevel hierarchical structures [26].

Since evaluation of criteria is modelled by ordinal preference relations over alternatives, in some cases, we have adapted or made use of preference aggregation methods in voting and social choice theory. There are broad number of voting rules, e.g. plurality, approval, Copeland's rule etc. Several properties are desired in a voting rule like Pareto's efficiency, monotonicity, Condorcet principle etc. The latter means that a voting rule should select the alternative that defeat every other alternative in pairwise comparisons [9]. Most of the well-known voting rules and their properties have already been discussed in social choice literature [23].

3 Framework

In this section, we describe our framework for the integration of moral values in the decision-making process. We are interested in evaluating the adherence or alignment of alternatives to a set of moral values. The alternatives have different signification according to the context, however they represent or refer to an action or a sequence of actions. We assume that all the alternatives are acceptable, i.e. they do not involve cases that violate a non derogatory right or ethical obligation. In order to consider moral values in the decision process, we view them as criteria in a hierarchical structure that can be decomposed to more basic measurable sub criteria. Evaluation is modeled as an ordinal preference relation over alternatives. We introduce multiple approaches with different logic for aggregating preferences that can take into account the priority of sub criteria. More precisely, our framework is designed to rank a set of given alternatives $\mathcal{A}$, $|\mathcal{A}| \geq 2$ based on their adherence to a set of moral values. A *hierarchical value setting* H over a set of alternative $\mathcal{A}$ is represented by the following tuple:

$$\langle \mathcal{A},\ \langle N,\ \mathbf{R}\rangle,\ \rho,\ \Psi\rangle \tag{1}$$

Criteria Hierarchy is a tree-like graph represented by the tuple $\langle N,\ \mathbf{R}\rangle$ where N is the set of nodes that represent criteria, $\mathbf{R} \subseteq N \times N$ is the child relation that assigns to each parent node its child nodes, in other words this relation associates each criteria with its sub criteria in a hierarchical structure. We denote the set of the sub criteria or the children of a node n by $r(n) = \{x|(n,x) \in \mathbf{R}, n \in N\}$. The root node in this structure is considered as the super criteria and is denoted by $\epsilon \in N$. The Tree-like hierarchical structure is shown in figure 1. Nodes at the bottom of the structure (leaves) represent evaluative criteria, i.e. criteria that can be evaluated based on the characteristics of the given alternatives. The set of leaf nodes or leaf criteria is represented by $L = \{n \in N|r(n) = \emptyset\}$.

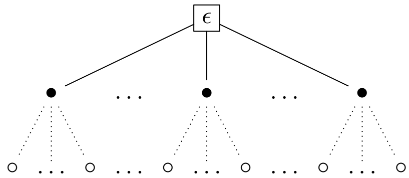

Figure 1. Criteria Hierarchy Diagram

Leaf Criteria Assessment ρ is the function to evaluate alternatives based on the bottom-level or leaf criteria. This evaluation is modeled as an ordinal preference over alternatives, and is represented as the following function.

$$\rho : L \longmapsto 2^{\mathcal{A}\times\mathcal{A}} \tag{2}$$

where for each $\rho(l),\ \forall l \in L$ is a preference relation, given as a complete preorder.

Aggregation functions $\Psi = \{\psi_n\}_{n\in N\setminus L}$ is a family of aggregators, used by each node to combine the preference of its children. To each non-leaf node n, we associate an aggregator ψ_n defined as a function from a vector of preference relations (whose size is the number of children of n) to a preference relation :

$$\psi_n : (2^{\mathcal{A}\times\mathcal{A}})^{|r(n)|} \longmapsto 2^{\mathcal{A}\times\mathcal{A}} \tag{3}$$

Output of A Hierarchical Value Setting . Given a hierarchical value setting H, by using its aggregation function of the preference relation of all its children, we can define the preference relation for each node n. We denote by $\mathbb{P}^H = \{P_n^H\}_{n \in N}$ the final set of preference for each node. It is defined inductively as follows:

$$\begin{aligned} P_n^H &= \rho(n) && \text{if } n \in L \\ &= \psi_n(P_{i_1}^H, \dots, P_{i_p}^H) && \text{if } r(n) = \{i_1, \dots, i_p\} \end{aligned}$$

Thus, given two alternatives a and b and a criteria n, $aP_n^H b$ means that a is preferred to b according to criteria n. The final evaluation of the model is then given by the preference of the node at the highest level, that is, P_ϵ^H. In the following, since we use a single hierarchical value setting, we drop the superscript H from the P_n^H notation.

4 Preference Aggregation

Here, we describe several aggregation methods that can be used in our hierarchical model to obtain the preference of a parent node based the preference of its children. In order to keep the generality of the problem, we consider a finite set of criteria $C = \{1, \dots, |C|\}$, such that each criterion $i \in C$ specifies a transitive and complete preference over the set of alternatives $P_i \subseteq D$, where $D = 2^{\mathcal{A} \times \mathcal{A}}$. We discuss, different rules for aggregating the ensemble of these preferences, noted by $\mathbb{P} = \langle P_1, \dots, P_{|C|} \rangle \in D^{|C|}$, according to the available information about the importance and priorities of criteria. The obtained preference by the aggregation rule is denoted by $P \subseteq D$. Each method is based on an assumption on the available information about the importance of criteria.

4.1 Voting Approaches

When all the criteria have an equal importance, their corresponding preference can be viewed as votes over alternatives. In such a case the aggregation rules in social choice and voting theory can be used to combine the preferences of the criteria. We denote this aggregator by $\psi_v : D^{|C|} \longmapsto D$ and the obtained preference by $P^v = \psi_v(\mathbb{P})$. There are general properties that are desired in voting rules, like Pareto's efficiency, monotonicity, etc. An important property in our case is the Condorcet principle, lack of this property may cause a cyclic preference and lead to loss of transitivity. We limit the rules to the ones that satisfy these properties, e.g. Copeland's rule, max-min etc. For example the Copeland's rule choose the alternative that win the most pairwise majority, for each $a, b \in \mathcal{A}$ let r_{ab} be defined as follows

$$r_{ab} = \begin{cases} 1 & |\{i \in C | aP_i b \wedge \neg bP_i a\}| > |\{i \in C | bP_i a \wedge \neg aP_i b\}| \\ \frac{1}{2} & |\{i \in C | aP_i b \wedge \neg bP_i a\}| = |\{i \in C | bP_i a \wedge \neg aP_i b\}| \\ 0 & |\{i \in C | aP_i b \wedge \neg bP_i a\}| < |\{i \in C | bP_i a \wedge \neg aP_i b\}| \end{cases} \tag{4}$$

Then, according to Copeland's rule an alternative wins in pairwise comparison if its overall score defined by the following relation is higher.

$$aP^v b \Leftrightarrow \sum_{c \in \mathcal{A}} r_{a,c} \geq \sum_{c \in \mathcal{A}} r_{b,c} \tag{5}$$

Note that this an example of a voting rule that have our desired properties, other rules that satisfy the Condorcet principle can be used interchangeably.

4.2 Voting with Dominant Voters

A possible situation is that there may be a group of criteria i.e. voters that dominate the others, meaning that the their vote has an absolute superiority over other voters. Here we suppose that the information about the relation between criteria is known and given by a total preorder $\succcurlyeq \subseteq C \times C$, which means that it is:

- Complete $(x \succcurlyeq y \; or \; y \succcurlyeq x), x \neq y, \forall x, y \in C$
- Transitive $(x \succcurlyeq y, y \succcurlyeq z) \rightarrow x \succcurlyeq z, \forall x, y, z \in C$

Given two criterion x and y, $x \succ y = x \succcurlyeq y \wedge \neg(y \succcurlyeq x)$ means that x dominates y and any other criterion less preferred to y, and $x \sim y = x \succcurlyeq y \wedge y \succcurlyeq x$ means x is as important as y. $\succcurlyeq$ can be seen as layers on C, such that the criteria inside a layer have equal importance and there is superiority among the criteria from different layers. This means C can be partitioned into $k \leq |C|$ components $C = \cup_{i \in \{1, \dots, k\}} C_i$ such that $\forall x, y \in C, x \succcurlyeq y \wedge y \succcurlyeq x \Leftrightarrow \exists i \in \{1, \dots, k\}, x, y \in C_k$. An example of $\succcurlyeq$ with 4 criteria has been depicted by a diagram in Figure 2, where each arrow shows a priority relation.

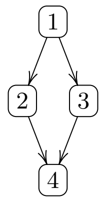

Figure 2. Example of $\succcurlyeq$ with 4 criteria

One way to aggregate the preference of criteria in this case is to aggregate the the votes at each layer based on a voting rule and select voting of the most dominant layer. More precisely, consider the aggregation function $\psi_{lay} : D^{|C|} \times C^2 \longmapsto D$ and its resulting preference $P^{lay} = \psi_{lay}(\mathbb{P}, \succcurlyeq)$. We define a total strict order $\succ^s \subseteq 2^C \times 2^C$ based on $\succcurlyeq$ such that $\forall X, Y \subseteq C, \; X \succcurlyeq^s Y \Leftrightarrow \forall x \in X, \forall y \in Y, x \succcurlyeq y \wedge \neg y \succcurlyeq x$. Then the aggregation method in this case can be formulated as follows, that is a combination of voting rule and lexicographic aggregation.

$$\begin{aligned} a \; P^{lay} \; b \Longleftrightarrow \; \exists \; i \in \{1, \dots, k\} \wedge a \; P_{C_i}^v \; b \; \wedge \\ (\forall \; j \in \{1, \dots, k\} \wedge C_j \succ^s C_i \; \Rightarrow a \; P_{C_j}^v \; b) \end{aligned} \tag{6}$$

Where $a \; P_X^v \; b = \psi_v(\mathbb{P}^X)$, $\mathbb{P}^X \subseteq D^{|X|}$ is the vector of the preference of the criteria in the set X, and ψ_v is an aggregation function based on a voting rule as discussed in the previous section.

4.3 General Approach

In previous case we assumed that the dominant group of voters have an absolute superiority over others. This means an alternative which gains the vote of the dominant group, always wins, regardless of the vote of other groups. in other words, let $X \subseteq C$ be a dominant group in C if it is preferred to any other group according to $\succ^s$, i.e we define $Dom(X, C) \Leftarrow \forall Y \subseteq C \wedge X \succ^s Y$ to be the set of groups that are dominated by X, then $\forall a, b \in \mathcal{A}$ we have the following property $\exists X \subseteq C \wedge aP_X^v b \wedge Dom(X, C) \Rightarrow aP^{lay} b$, which means if X is the dominant group in C then its vote determines the final preference. This is because of assuming an absolute superiority among certain groups. However, in general we can assume more comprehensive rules for aggregation other than superiority.

Here we propose a method for aggregating votes by comparing directly the set of voters that favor each alternative using a transitive preference $\succeq \subseteq 2^C \times 2^C$. we denote the set of criteria that support an alternative a by $C_a = \{i \in C | a \ P_i \ x, \exists x \in \mathcal{A}\}$. The advantage of a preference relation on 2^C is two fold. First, taking into account available knowledge between any subset of criteria coming form a known source e.g. elicited preferences. Second, it allows incorporating more specific aggregation behavior by inducing certain properties. however $\succeq$ may not always be complete. The aggregation function in this case is $\psi_{par} : D^{|C|} \times (2^C \times 2^C) \longmapsto D$ that aggregates votes based on $\succeq$. We denote the aggregated preference by $P^{par} = \psi_{par}(\mathbb{P}, \succeq)$ defined as:

$$a \ P^{par} \ b \Longleftrightarrow C_a \succeq \ C_b \tag{7}$$

This method allows building specific aggregation rules when some preferences are known and certain aggregation properties are desired. in order to describe the process consider a simple example with 3 criteria i.e. $C = \{1, 2, 3\}$. we suppose the following items are assumed about $\succeq$ which include known preferences and properties that may be desirable depending on the context

- 1) Known preferences: $\{1\} \succeq \{2\}$, $\{2\} \succeq \{3\}$ and $\{2, 3\} \succeq \{1\}$ the known preferences imply that there is a linear order among single criteria but the two lower order criteria beat the first one. This can be shown by Figure 3.
- 2) A given set of criteria is preferred to any of its subsets i.e. $\forall \ X, Y \subseteq C, X \subseteq Y \Rightarrow Y \succeq X$. This property indicates that the criteria have a positive signification.
- 3) The preference among two given set of criteria is independent of the common elements among them i.e. $\forall \ X, Y \subseteq C, \ X \setminus Y \succeq Y \setminus X \Rightarrow X \succeq Y$.

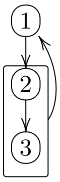

Figure 3. Known preferences in $\succeq$ with 3 criteria

Figure 4 shows how $\succeq$ would be effected by inducing the mentioned elements step by step. Note that in this example we have finally acheved a complete order, but this may not be the case for higher number of criteria.

5 Case Study: Ranking of Recommender Systems

In this section, we show how our framework can be used for ethical evaluation regarding moral values. The set of responsible AI values is not fixed and not all of them apply to any given problem. On the other hand the interpretation of moral values and the criteria by which they are assessed are highly dependent to the context. However, most of the current applications of AI fall into a limited number categories where the set of applicable responsible values and their evaluative criteria are known. The collection of this knowledge can be represented by an ontology and serve as a complementary tool for the current framework which we let for future development. We describe the evaluation procedure on a use case model of ranking recommender systems, which concerns privacy, fairness and performance of the systems. Our framework in such context provide an explicit evaluation process that is explainable in terms of responsible AI values by indicating what moral values it takes into account and how it prioritizes them. This problem is modelled in ASP to show the reasoning capabilities, of our framework, which is essential in ethical decision-making. The codes are available online [1].

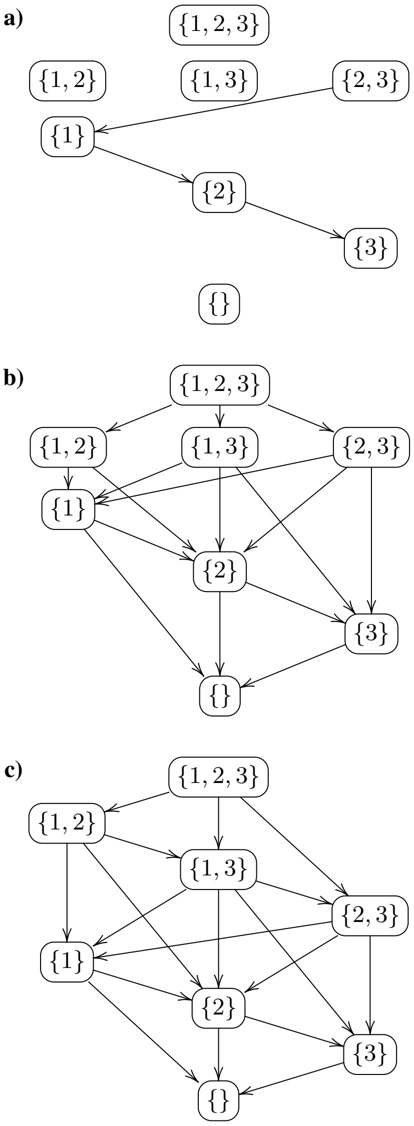

Figure 4. The step by step induction Process, (sets without an arrow are not comparable) **a)** Induction of known preferences, **b)** Induction of property 2, and **c)** induction of property 3

5.1 Use Case Description

A company with an online employment platform collaborates with multiple partners to provide job recommendations for its users. The Company's partners change regularly and/or update their recommender systems very often. The company is using an automatic ranking process to select the best system by taking into account responsible AI values. We show the evaluation process of our framework

[1] https://anonymous.4open.science/r/responsibleAi-A852

on a simplified version of this problem for ranking a set of 3 recommender systems i.e. alternatives `{sys1, sys2 ,sys3}`. Each alternative is represented in ASP using the predicate `alt/1`.

```
alt(sys1).
alt(sys2).
alt(sys3).
```

5.2 Criteria Hierarchy

Here we describe what moral values are taken into account and by which criteria the alternatives are evaluated. Ethical evaluation in this context concerns with *i) Privacy* of the users since these systems process personal data. *ii) Fairness* of the system with regard to jobs and users, and *iii) Performance* of the system in providing the fittest recommendations. Thus, the set of the moral values in this case to be integrated in the evaluation framework is $\{Privacy,\ Fairness,\ Performance\}$. Next, is to identify how can we compare the alternatives with regard to these moral values. In other words, what are the sub-criteria of the given moral values that can evaluate alternatives.

- **Privacy** At this stage of AI privacy is one most concerning issues. we propose some criteria in this case that can be used to evaluate the alternatives according to their accordance with privacy.
  - **Data Minimisation** limiting the access to data categories is an important and obvious criterion for a preserving data subjects privacy. A system which uses less categories of data ranked higher by this criteria
  - **Data Sensitivity** Another important criterion is the number of sensitive categories of data e.g. political, racial, etc. Processing these type of data entails a higher ethical risk to privacy [2].
  - **Scale and Complexity** This is another important criteria related to privacy. Using large scale processing i.e. big data with multiple unknown sources of personal data increase the risk to privacy [2]. according to this criteria small scale processing get a higher rank than large scale big data processing
- **Fairness** in general has a broad meaning, however in this context it means having no bias toward any particular group of items or users [30]. User fairness can also be decomposed to gender fairness and racial fairness. Fairness can be measured by a metric called *statistical parity* which is the difference in the ratio of favorable recommendations between two groups. A recommender system that has a lower statistical parity among monitored groups is ranked higher by fairness criteria. The sub criteria of fairness in this case are:
  - **Item Fairness**
  - **User Fairness**
    - **Gender Fairness**
    - **Racial Fairness**
- **Performance** represents the intrinsic value of the recommender system that has been designed to serve for. The notion of performance has various significations according to the context, here it represents how fit the recommendations are and if the users are reacting positively to the generated recommendations. A system that have a better performance score has more adherence to this value. Here we suppose that that performance represent a single criterion without any sub criteria.

The value composition diagram in this use case is given by Figure 5. We use the predicate `child/2` in ASP in order to represent the criteria hierarchy.

```
child(root,privacy).
child(root,fairness).
child(root,performance).

child(privacy, sensitivity).
child(privacy, minimization).
child(privacy, scaleComplexity).
child(fairness, item_fairness).
child(fairness, user_fairness).
child(user, racial_fairness).
child(user, gender_fairness).
```

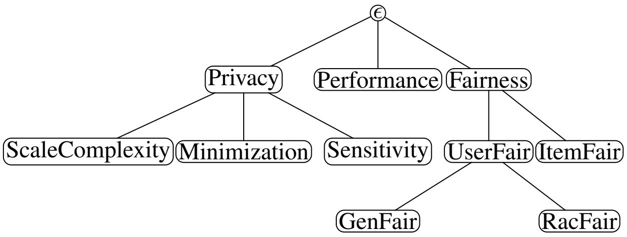

Figure 5. Criteria Hierarchy Diagram

5.3 Leaf Criteria Assessment

In order to apply our framework in this case study, we need to assess the preference of the leaf criteria over alternatives according to their characteristics. We should first represent these characteristics or the knowledge about the recommender systems and their relevant characteristic that can be used for ethical evaluations. These characteristics include i) the knowledge about the required categories of data, ii) the performance metric of each system, iii) The scale and complexity of the underlying algorithm. And the statistical parity for iv) gender fairness, v) racial equality as well as vi) the statistical parity for item fairness. The knowledge about the available systems and their mentioned characteristics are shown in table 1

Characteristics	**Sys1**	**Sys2**	**Sys3**
Data Categories	dataHabit, activity, interests	interests, politicalBelief	activity, interests
Performance metric	30%	25%	20%
Process Type	large scale	larg scale	small scale
Gender parity	0,1	0,2	0,3
Racial Parity	0,2	0,4	0,1
Item Parity	0,3	0,6	0,4

Table 1. Recommender Algorithms' Characteristics

This knowledge has been represented in ASP using the predicate `has/3`. Note that since Clingo has difficulties with grounding float numbers we represent them using integers. for example:

```
has(sys1,requiredData,clickHabit ).
has(sys1,requiredData,activity ).
has(sys2,requiredData,interests ).
...
```

```
    has(sys1, perfMetric, 30 ).
    has(sys2, perfMetric, 25 ).
    ...
    has(sys1, gender_parity, 1 ).
    has(sys2, racial_parity, 4 ).
    has(sys3, item_parity, 4 ).
    ...
```

Having these characteristics, the assessing functions for the leaf criteria evaluates the alternatives in a pairwise manner based on the relevant characteristics of each alternative. The preference of the leaf criteria is represented by the predicate `pref/3`. We suppose that the evaluated preferences are complete and transitive, and we use the ranking of alternatives to represent them as in Table 2.

Leaf Criteria	Sys1	Sys2	Sys3
Minimization	2	1	1
Sensibility	1	2	1
ScaleComplexity	2	2	1
Performance	1	2	3
Gender fairness	1	2	3
Racial fairness	2	3	1
Item fairness	1	3	2

Table 2. Recommender Systems' Leaf Ranking

```
% Minimisation
has(Alt,nbData, N):-
    N = #count{ Data : has(Alt, requiredData, Data)},
    alt(Alt).

pref(minimization,Alt1,Alt2):-
    has(Alt1, nbData, N1),
    has(Alt2, nbData, N2),
    N2>=N1.

%Sensitivity
has(Alt, nbSensitiveData, N):-
    N = #count{ Data : has(Alt, requiredData, Data),
    has(Data, category, sensitiveData)},
    alt(Alt).
pref(sensitivity, Alt1, Alt2):-
    has(Alt1, nbSensitiveData, N1),
    has(Alt2,nbSensitiveData, N2),
    N2>=N1.

%Scale and Complexity
rankAux(largeScale, 2).
rankAux(smallScale, 1).
pref(scaleComplexity,Alt1,Alt2):-
    has(Alt1, processType, Type1),
     has(Alt2, processType, Type2),
     rankAux(Type1, R1),
     rankAux(Type2, R2),
     R2>=R1, alt(Alt1), alt(Alt2).

% Performance
pref(performance, Alt1, Alt2):-
    has(Alt1, perfMetric, P1),
    has(Alt2, perfMetric, P2),
    P1>=P2, alt(Alt1), alt(Alt2).

% Item Fairness
pref(item_fairness,Alt1,Alt2):-
    has(Alt1, item_parity,P1),
    has(Alt2, item_parity,P2),
    P1<=P2, alt(Alt1), alt(Alt2).

%Item Fairness
pref(racial_fairness,Alt1,Alt2):-
    has(Alt1, racial_parity,P1),
    has(Alt2, racial_parity,P2),
    P1<=P2, alt(Alt1), alt(Alt2).

%Gender Fairness
pref(gender_fairness,Alt1,Alt2):-
    has(Alt1, gender_parity,P1),
    has(Alt2, gender_parity,P2),
    P1<=P2, alt(Alt1), alt(Alt2).
```

5.4 *Aggregation*

Here, we describe the aggregation process for every parent node given a preference over their children We suppose these preferences are complete and use the layered aggregation function in order to obtain the preference of the parent nodes. The root node is the top-most node that represent the final decision criteria. Here, we consider multiple priority settings for the root node to show the functionality of the aggregator and expressivity of our model, as in table 3.

Parent node	Children priorities Criteria	Sys1	Sys2	Sys3
Privacy	Sensibility $\succcurlyeq$ Minimization $\succcurlyeq$ ScaleComplexity	1	2	1
User Fairness	genderFairness $\sim$ racialFairness	1	2	2
Fairness	userFairness $\succcurlyeq$ itemFairness	1	2	2
root	Case1: Fairness $\sim$ Privacy $\sim$ Performance	1	2	2
root	Case2: Fairness $\succcurlyeq$ Privacy $\succcurlyeq$ Performance	1	2	2
root	Case3: Fairness $\sim$ Privacy $\succcurlyeq$ Performance	1	2	1
root	Case4: Fairness $\sim$ Performance $\succcurlyeq$ Privacy	1	1	2
root	Case5: Performance $\succcurlyeq$ Fairness $\succcurlyeq$ Privacy	1	2	3

Table 3. Parent Criteria Evaluations

We represent the priority of sub values using the predicate `childPref/3` in ASP as shown below.

```
childPref(privacy, sensitivity, minimization).
childPref(privacy, minimization, scaleComplexity).

childPref(fairness, user_fairness, item_fairness).
childPref(user_fairness,racial_fairness,gender_fairness).

childPref(root, fairness, privacy).
childPref(root, privacy, performance).
```

The Layered Aggregation function 6 in ASP has been implemented in the following way.

```
prefLay(Node, Alt1, Alt2):-
    childrenLayers(Node,Layer),
    pVote(Node, Layer, Alt1, Alt2),
    is_dominant(Node, Layer, Alt1, Alt2).

is_dominant(Node, Layer,Alt1,Alt2):-
    not is_dominated(Node, Layer,Alt1,Alt2),
    childrenLayers(Node,Layer),
    pVote(Node, Layer, Alt1, Alt2).

is_dominated(Node, Layer, Alt1,Alt2):-
    childrenLayers(Node,Layer),
    pVote(Node, Layer, Alt1, Alt2),
    superiorThan(Node, Layer1, Layer),
    not pVote(Node, Layer1, Alt1,Alt2).

superiorThan(Node, Layer1, Layer2):-
    childrenLayers(Node, Layer1),
    childrenLayers(Node, Layer2),
    not inferiorThan(Node, Layer1, Layer2).

inferiorThan(Node, Layer1, Layer2):-
    childrenLayers(Node, Layer1),
    childrenLayers(Node, Layer2),
    belongs(Node, Child1 , Layer1),
    belongs(Node, Child2 , Layer2),
    not childPref(Node, Child1, Child2).

childrenLayers(Node, Layer):-
    belongs(Node, _ , Layer).

belongs(Node, Child , Layer):-
    Layer = #count{ Child1 :
        childPref(Node , Child, Child1)},
    child(Node, Child).
```

The predicate `pVote/4` in the code above corresponds to a voting rule. In our case we have used Copeland's rule in order to aggregate votes in each layer. The rule is translated in ASP in the following way that is an implementation of 5:

```
pVote(Node, Layer, Alt1, Alt2):-
    plural(Node, Layer),
    copeland_score( Node, Layer, Alt1, S1),
    copeland_score( Node, Layer, Alt2, S2),
    S1>S2.
```

```
copeland_score( Node, Layer, Alt, S):-
    S= #sum{ S1:
        nb_pairwise_wins( Node, Layer, Alt, Alt1, N1),
        nb_pairwise_ties( Node, Layer, Alt, Alt1, N2),
        S1 = N1*2+N2,
        alt(Alt1)},
    childrenLayers(Node, Layer), alt(Alt) .

nb_pairwise_ties( Node, Layer, Alt1, Alt2 , N):-
    N= #count{ N1 :
        nb_strict_voters(Node, Layer, Alt1, Alt2 , N1),
        nb_strict_voters(Node, Layer, Alt2, Alt1 , N2),
        belongs(Node, Child, Layer), N1=N2},
    childrenLayers(Node,Layer), alt(Alt1), alt(Alt2).

nb_pairwise_wins( Node, Layer, Alt1, Alt2, N):-
    N= #count{ N1 :
        nb_strict_voters(Node, Layer, Alt1, Alt2 , N1),
        nb_strict_voters(Node, Layer, Alt2, Alt1 , N2),
        belongs(Node, Child, Layer), N1>N2},
    childrenLayers(Node, Layer), alt(Alt1), alt(Alt2).

nb_strict_voters(Node, Layer, Alt1, Alt2, N):-
    N = #count{ Child :
        pref(Child, Alt1, Alt2),
        not pref(Child , Alt2, Alt1),
        belongs(Node, Child, Layer) },
    childrenLayers(Node, Layer), alt(Alt1), alt(Alt2).
```

The aggregated results for each parent node is shown in Table 3. For the root node we consider several cases in order to show the aggregation result given different orders on the moral values. This shows our framework capacity to take into account subjective preferences on moral values. In some cases the children priorities are different however the ranking of alternatives is the same. For example when all the root children are considered equal (case 1) $Sys1$ is selected as the best recommender, and when there is a superiority between moral values $Sys1$ is selected again. In such case even if the rankings are equal the logic behind each one is different. In case 1 the rankings are obtained by collective voting of all moral values but in case 2 the rankings of the dominant voter, i.e. fairness, has selected over others. by using this framework, we can obtain an ethical evaluation in terms of moral values. Our framework also considers the subjectivity of the ethical decision making by considering various value systems given as a priority relation among values. For example In a health care scenario we may wish to prioritize performance over privacy but in a marketing context we tend to prioritize privacy over performance.

6 Conclusions

The proposed model for ethical evaluation provides an explicit representation of ethical considerations in terms of responsible AI values, clarifies the interpretation of these values and can take their importance into account by prioritizing them. An advantage of our model is that the rankings are obtained by logical aggregation rules through a reasoning process which is essential in ethical decision making. our framework can also serve as a utility function for modelling consequentialist theories. One of the future work is designing an ontology to collect the principles and prescriptions in AI ethics guidelines which can cover a broad range of responsible-AI values

References

[1] Ethics guidelines for trustworthy ai.

[2] Ethics and data protection, 2021.

[3] Michael Anderson, Susan Leigh Anderson, and Chris Armen, 'Towards machine ethics', in *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA*, (2004).

[4] Michael Anderson, Susan Leigh Anderson, and Chris Armen, 'MedEthEx: a prototype medical ethics advisor', in *Proceedings of the national conference on artificial intelligence*, volume 21, p. 1759. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, (2006).

[5] Martin Aruldoss, T Miranda Lakshmi, and V Prasanna Venkatesan, 'A survey on multi criteria decision making methods and its applications', *American Journal of Information Systems*, **1**(1), 31–43, (2013).

[6] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia, 'A declarative modular framework for representing and applying ethical principles', in *16th Conference on Autonomous Agents and MultiAgent Systems*, (2017).

[7] Craig Boutilier, Ronen I Brafman, Carmel Domshlak, Holger H Hoos, and David Poole, 'Cp-nets: A tool for representing and reasoning withconditional ceteris paribus preference statements', *Journal of artificial intelligence research*, **21**, 135–191, (2004).

[8] Steven J Brams, *Approval voting*, Springer, 2004.

[9] Felix Brandt, Vincent Conitzer, and Ulle Endriss, 'Computational social choice', *Multiagent systems*, **2**, 213–284, (2012).

[10] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier, 'Ethical Judgment of Agents' Behaviors in Multi-Agent Systems.', in *AAMAS*, pp. 1106–1114, (2016).

[11] Marie Jean Antoine Nicolas de Caritat and Marquis De Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, 1785.

[12] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub, 'Clingo= asp+ control: Preliminary report', *arXiv preprint arXiv:1405.3694*, (2014).

[13] Salvatore Greco, Jose Figueira, and Matthias Ehrgott, *Multiple criteria decision analysis*, volume 37, Springer, 2016.

[14] Martin Jedwabny, Pierre Bisquert, and Madalina Croitoru, 'Generating preferred plans with ethical features', in *Florida Artificial Intelligence Research Society*, volume 34, (2021).

[15] Anna Jobin, Marcello Ienca, and Effy Vayena, 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, **1**(9), 389–399, (2019).

[16] John G Kemeny and LJ Snell, 'Preference ranking: an axiomatic approach', *Mathematical models in the social sciences*, 9–23, (1962).

[17] Vladimir Lifschitz, *Answer set programming*, Springer Heidelberg, 2019.

[18] Raynaldio Limarga, Maurice Pagnucco, Yang Song, and Abhaya Nayak, 'Non-monotonic reasoning for machine ethics with situation calculus', in *AI 2020: Advances in Artificial Intelligence: 33rd Australasian Joint Conference, AI 2020, Canberra, ACT, Australia, November 29–30, 2020, Proceedings 33*, pp. 203–215. Springer International Publishing, (2020).

[19] Felix Lindner, Robert Mattmüller, and Bernhard Nebel, 'Evaluation of the moral permissibility of action plans', *Artificial Intelligence*, **287**, 103350, (2020).

[20] Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K Brent Venable, 'Value alignment via tractable preference distance', in *Artificial Intelligence Safety and Security*, 249–261, Chapman and Hall/CRC, (2018).

[21] Elinor Mason, 'Value pluralism', (2006).

[22] Brent Mittelstadt, 'Principles alone cannot guarantee ethical AI', *Nature machine intelligence*, **1**(11), 501–507, (2019).

[23] Hannu Nurmi, 'Voting systems for social choice', in *Handbook of Group Decision and Negotiation*, 167–182, Springer, (2010).

[24] Zdzisaw Pawlak and Roman Sowinski, 'Rough set approach to multi-attribute decision analysis', *European journal of Operational research*, **72**(3), 443–459, (1994).

[25] Catharina Rudschies, Ingrid Schneider, and Judith Simon, 'Value pluralism in the AI ethics debate–different actors, different priorities', *The International Review of Information Ethics*, **29**, (2020).

[26] Thomas L Saaty, Luis G Vargas, et al., *Decision making with the analytic network process*, volume 282, Springer, 2006.

[27] Mark Schroeder, 'Value theory', (2008).

[28] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein, 'Implementations in machine ethics: A survey', *ACM Computing Surveys (CSUR)*, **53**(6), 1–38, (2020).

[29] Luis G Vargas, 'An overview of the analytic hierarchy process and its applications', *European journal of operational research*, **48**(1), 2–8, (1990).

[30] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma, 'A survey on the fairness of recommender systems', *ACM Transactions on Information Systems*, **41**(3), 1–43, (2023).

[31] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave, 'The role and limits of principles in AI ethics: towards a focus on tensions', in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200, (2019).

[32] Constantin Zopounidis and Michael Doumpos, 'Multicriteria classification and sorting methods: A literature review', *European Journal of Operational Research*, **138**(2), 229–246, (2002).