Aggregating value systems for decision support

Lera-Leri, Roger X.; Liscio, Enrico; Bistaffa, Filippo; Jonker, Catholijn M.; Lopez-Sanchez, Maite; Murukannaiah, Pradeep K.; Rodriguez-Aguilar, Juan A.; Salas-Molina, Francisco

DOI
10.1016/j.knosys.2024.111453

Publication date
2024

Document Version
Final published version

Published in
Knowledge-Based Systems

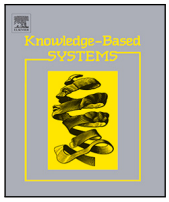
Citation (APA)
Lera-Leri, R. X., Liscio, E., Bistaffa, F., Jonker, C. M., Lopez-Sanchez, M., Murukannaiah, P. K., Rodriguez-Aguilar, J. A., & Salas-Molina, F. (2024). Aggregating value systems for decision support. *Knowledge-Based Systems*, *287*, Article 111453. https://doi.org/10.1016/j.knosys.2024.111453

Knowledge-Based Systems 287 (2024) 111453

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Aggregating value systems for decision support

Roger X. Lera-Leri [a,*], Enrico Liscio [b], Filippo Bistaffa [a], Catholijn M. Jonker [b], Maite Lopez-Sanchez [c], Pradeep K. Murukannaiah [b], Juan A. Rodriguez-Aguilar [a], Francisco Salas-Molina [d]

[a] Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Spain
[b] Delft University of Technology, Delft, The Nederlands
[c] Universitat de Barcelona, Barcelona, Spain
[d] Universitat Politècnica de València, Alcoi, Spain

ARTICLE INFO

Keywords:
AI & ethics
Value systems
Optimisation

ABSTRACT

We adopt an emerging and prominent vision of human-centred Artificial Intelligence that requires building trustworthy intelligent systems. Such systems should be capable of dealing with the challenges of an interconnected, globalised world by handling plurality and by abiding by human values. Within this vision, pluralistic value alignment is a core problem for AI– that is, the challenge of creating AI systems that align with a set of diverse individual value systems. So far, most literature on value alignment has considered alignment to a single value system. To address this research gap, we propose a novel method for estimating and aggregating multiple individual value systems. We rely on recent results in the social choice literature and formalise the value system aggregation problem as an optimisation problem. We then cast this problem as an ℓ_p*-regression problem.* Doing so provides a principled and general theoretical framework to model and solve the aggregation problem. Our aggregation method allows us to consider a range of *ethical principles*, from utilitarian (maximum utility) to egalitarian (maximum fairness). We illustrate the aggregation of value systems by considering real-world data from two case studies: the Participatory Value Evaluation process and the European Values Study. Our experimental evaluation shows how different consensus value systems can be obtained depending on the ethical principle of choice, leading to practical insights for a decision-maker on how to perform value system aggregation.

1. Introduction

The vision of human-centred Artificial Intelligence (AI) has spurred research on trustworthy, ethical AI that enhances human capabilities and empowers citizens and society to deal with the globalised world's challenges effectively. Thus, developing trustworthy AI [1] that abides by human values is a primary AI concern, as explicitly stated by the European Commission's Ethics Guidelines for Trustworthy AI [2], the Artificial Intelligence Act [3], and the IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems [4]. Within this vision, a core problem is AI value alignment, which aims to ensure that AI is properly aligned with human values [5]. Designing an AI agent to align with human values means that the agent does "what it morally ought to do, as defined by the individual or society"[1] [6].

The problem of value alignment has spurred research on different aspects of the challenge such as formalising the value alignment problem [7], identifying relevant values [8], value-sensitive design [9–11], learning value-aligned behaviours [12–18], reasoning about values to act ethically [19–21], and aligning norms with values [22–24]. A core component of the different value alignment endeavours is the concept of a *value system*, which models how an entity (e.g., an individual or an organisation) interprets and prioritises values. A common assumption in most state-of-the-art research on value alignment is that an AI system must align with *one* value system, be it that of an individual or a society. However, as Gabriel [6] argues, by following Rawls, humans may hold various reasonable but contrasting beliefs about values. That is, we live in a pluralistic world where people hold different value systems. Designing an AI system that aligns with a group of people with different value systems poses the *pluralistic value alignment* problem [6]. This is the case, for instance, when making policy decisions that align with stakeholders having a variety of value systems (e.g., when deciding over water governance [25,26]) or when designing human-agent teams

* Corresponding author.
E-mail address: rlera@iiia.csic.es (R.X. Lera-Leri).

[1] Therefore, hereafter we shall refer to human values as *moral values*, or simply as *values* for shorter when there is no risk to incur in confusion.

https://doi.org/10.1016/j.knosys.2024.111453
Received 22 December 2022; Received in revised form 24 January 2024; Accepted 26 January 2024
Available online 1 February 2024

involving humans with differing ethical perspectives (the most apparent examples exist in the medical field, where teams are constantly tasked with scenarios that require ethical consideration) [27].

In this paper, we address the following research question: how to determine the value system(s) that an AI system should align with, considering value diversity? More precisely, we make headway in the pluralistic value alignment problem by addressing the aggregation of different value systems to yield a *consensus* value system. We assume that the value systems to aggregate belong to any form of individual entities, be they citizens, stakeholders, teammates, countries, or even supra-national entities. To succeed in this endeavour, we identify three major challenges.

First, as noted by Mittelstadt [28], existing ethical codes are rather abstract about moral values, hence not specific enough to guide actions. This is also typically the case in the AI literature on value systems (e.g., [23,29–31]). Therefore, it is necessary to precisely identify the key elements that a value system must include. Importantly, this effort must consider that: (i) value systems are *contextual* [8,32–34], i.e., the way in which we reason about and prioritise values is influenced by contextual factors such as actors and actions [35–37]; and (ii) individuals may ascribe different *interpretations* to the same value system [38]. Second, we need to *estimate* or characterise the value system of each individual [39]. Third, from a social choice perspective, value systems can be aggregated following different *ethical principles* (e.g., utilitarian or egalitarian). Therefore, we need a principled and general aggregation method that allows us to set the ethical principle of choice.

We approach the challenge of pluralistic value alignment by studying two real-world cases. On the one hand, we consider the data from a Participatory Value Evaluation (PVE) [40] process conducted in 2020 to gauge the opinion of the residents of a municipality in the Netherlands over energy policies [41]. From this data, we estimate the value system of each citizen who participated in the PVE. On the other hand, we consider data from the European Values Study (EVS) [42], an extensive survey conducted on European citizens to characterise the value systems of European countries.

We build on preliminary works on value aggregation [38] and value estimation [43] to make the following novel contributions:

- We extend the definition of value system to account for *contextuality* and *pluralism*. We show how the novel definition can be employed to represent individual value systems in the two case studies that we analyse.
- Based on the social choice (distance) functions from González-Pachón and Romero [44], we formalise the problem of aggregating different value systems following a given ethical principle (e.g., utilitarian or egalitarian), and we cast it as a two-step optimisation problem to obtain: (i) the aggregation of value interpretations from individuals, and (ii) the aggregation of the preferences of individuals over moral values.
- We show that the problem of computing the consensus value system can be cast as an ℓ_p*-regression problem* [45] (also called *norm approximation problem* [46]). By doing so, we provide a general theoretical framework that allows us to solve the above-mentioned problem for a range of ethical principles – from utilitarian (maximum utility) to egalitarian (maximum fairness) – in a scalable and reliable way, thanks to recent results in the machine learning literature [45].
- We illustrate our value aggregation approach with real-world data from the PVE [41] and the EVS [42]. We conduct a thorough empirical evaluation that : (i) shows the impact of choosing different ethical principles on the resulting consensus value system; (ii) characterises the space of ethical principles to determine whether a given ethical principle produces a consensus leaning towards the utilitarian or the egalitarian ethical principle; and (iii) quantifies the price to pay when moving away from the majority to include the minority. The aggregation results differ in the two case studies since they handle different data. However, we delineate the common trends we observe. Finally, we provide practical insights for decision-makers concerned with obtaining a consensus on different value systems according to an ethical principle of choice.

In summary, we provide both the computational means and the guidelines for a decision-maker to conduct the principled aggregation of value systems.

Organisation. Section 2 provides background on the distance functions that we require from the social choice literature. Section 3 identifies the key elements of a value system and introduces the value systems in our two case studies. Section 4 formalises our aggregation problem and Section 5 shows that it can be cast as an ℓ_p-regression problem that can be solved as described in Section 5.1. Section 6 reports on our empirical findings and provides guidelines for the decision-makers in charge of the value system aggregation. Finally, Section 7 draws conclusions and sets paths to future research.

2. Preliminaries: Distance functions

As mentioned above, the main goal of this paper is to aggregate the values and preferences on values (i.e., the value systems) of different individuals to yield a *consensus* value system. The basic problem in designing socially optimal decisions is aggregating individual preferences on multiple alternatives into a collective preference representing a consensus. This problem is the core of disciplines such as social choice (e.g., [47,48]), multi-expert decision-making (e.g., [49,50]), or group decision-making (e.g., [51–54]). Therefore, many procedures for undertaking this aggregation task have been proposed in the literature. However, methods for aggregating preferences holding "good" theoretical properties are needed to facilitate the acceptance of the resulting consensus by the group of individuals involved in the decision [44]. Designing aggregation functions that exhibit good *social choice* properties (e.g., unanimity, anonymity, non-dictatorship) is a major goal in the social choice literature [47]. For this reason, to tackle the aggregation problem that we pose in this paper, we resort to the tools in the social choice literature.

This section provides a background on the social choice functions that we employ. We borrow from existing work [44,55–57], which define a generator of social choice functions (as a p-parameterised distance function) to obtain a consensus in a society of individuals. The choice of this generator of social choice functions is motivated by several reasons. First, it is well-founded on the social choice literature and multi-criteria decision-making literature, following earlier work by Cook [58–61] and Yu's p-metric distance [62,63] respectively. Second, the literature (e.g., [55,64]) has already studied the social choice properties underlying the compromise consensuses derived from the selected social choice function (e.g., neutrality, monotonicity, anonymity). Third, the work in [56] shows that our social choice function allows us to solve a wide range of aggregation situations (involving both ordinal and cardinal preferences of individuals). Finally, the works in [44,55] offer interpretations of the consensus that the social choice function produces. This includes a study (limited to the cases of $p = 1$ and $p = \infty$) of the *ethical* interpretation of the resulting consensus within a context of social choice. This means that our generator produces social functions that vary depending on ethical principles (e.g., egalitarian, utilitarian, equity).[2] Such ethical interpretations provide us with foundations to analyse consensuses.

[2] In this paper, we do not consider the consensus computed considering the principle of equity (i.e., the so-called Marxian solution) since, as noted by the authors of [44], it often results in an over-constrained optimisation problem that yields no solution.

Thus, one of the major benefits of our generator of social choice functions is that it allows decision-makers to compute a consensus following different *ethical principles* according to different values of the parameterised distance function.

The general setting of González-Pachón and Romero [55] considers a society of n members ($i = 1, \dots, n$). Each member of the society provides judgement on m objects ($j = 1, \dots, m$), which can be candidates, criteria, alternatives, etc. In the case of aggregating value systems, the objects represent moral values. As argued in Section 3, rankings are typically used to describe value systems. Thus, we consider that each member of the society ranks their moral value preferences. Furthermore:

- w_i is the weight (social influence) of the ith member.
- $R_i[j]$ is the rank position provided by the ith member of the society for the jth object (in our case, with the jth value) within the ranking.
- $R_S[j]$ is the *consensus* position assigned by the society as a whole to the jth object (in our case, with the jth value). R_S is the unknown consensus ranking that we seek to obtain.
- p is a metric parameter (i.e., an integer ≥ 1) that determines the *ethical principle* used to compute the consensus, in accordance with the terminology established in the social choice literature [44,55–57].

From the previous definitions, a generator of social choice functions based on the weighted Minkowski p-metric distance function (U_p) is introduced and described in several works [44,55–57] as:

$$U_p = \left[\sum_{i=1}^{n} \sum_{j=1}^{m} w_i |R_i[j] - R_S[j]|^p \right]^{1/p}. \tag{1}$$

Given the distance function U_p and a value of p, the goal is to find the *consensus* values of $R_S[j]$ that minimise the deviation between the judgements provided by the members of the society (data of the problem) and the consensus (the unknown).

In addition, González-Pachón and Romero [44] modify Eq. (1) to aggregate more complex objects besides rankings. In this case, they propose a distance function to aggregate a 2-dimensional vector such that

$$U_p = \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1, k\neq j}^{m} w_i |R_i[j,k] - R_S[j,k]|^p \right]^{1/p}, \tag{2}$$

where $R_i[j,k]$ is the judgement value provided by the ith member of the society when comparing the jth and the kth object. It is relevant that González-Pachón and Romero [56] shows how the social choice function in Eq. (2) can handle a wide range of aggregation situations. This includes aggregations when information about preferences is: ordinal and complete (complete rankings); ordinal and partial (partial rankings) ; and cardinal and complete (pairwise comparisons). Thus, when using Boolean values in the preference (R) matrix, we can express ordinal and qualitative preferences, whereas we can also express graded quantitative preferences through cardinal values. Furthermore, as argued in [44], the information in the preference matrix can be complete or incomplete.

From the U_p distance function, González-Pachón and Romero [44] derive two cases of interest. First, by setting $p = 1$, the general distance in Eq. (1) yields

$$U_B = \left[\sum_{i=1}^{n} \sum_{j=1}^{m} w_i |R_i[j] - R_S[j]| \right]. \tag{3}$$

The consensus that minimises U_B provides the social optimum from the point of view of the majority, i.e., the *utilitarian* solution (or Benthamite solution [65]) that maximises the total welfare.

By setting $p = \infty$, the distance function in Eq. (1) yields

$$U_R = \max_{i,j} \left[|R_i[j] - R_S[j]| \right]. \tag{4}$$

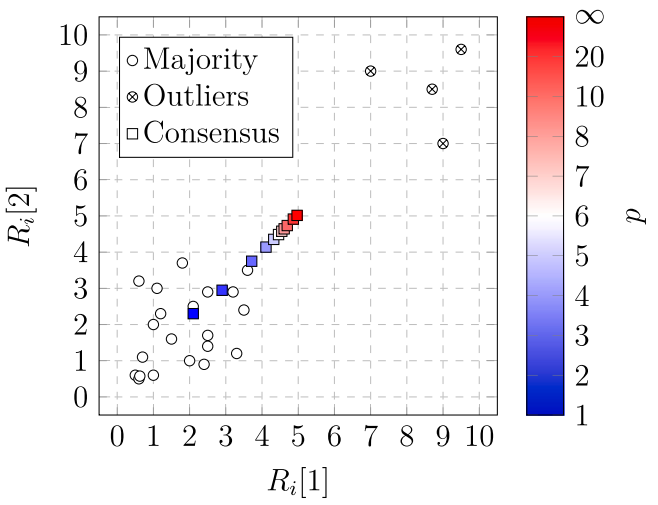


Fig. 1. The impact of the ethical principle p on the computed consensus of a set of individuals' judgement of two objects: $R_i[1]$ and $R_i[2]$. Circles show the individuals' judgement and squares represent the consensus computed. Squares are filled with the colour of the ethical principle used to compute the consensus.

Eq. (4) represents the Chebyshev distance, which is equivalent to the weighted Minkowski distance for $p = \infty$. In this case, finding the consensus implies minimising the disagreement of the member of the society most displaced with respect to the majority solution defined by the utilitarian case above (Eq. (3)). This solution is *egalitarian* [66] since it represents the social optimum from the point of view of the minority (from the perspective of the worst-off member of the society according to the Rawls' principle [67]), leading to the point of maximum fairness. Note that, when considering the limit case $p \to \infty$, the weighting scheme in Eq. (1) vanishes, hence the weight w_i does not appear in Eq. (4).

In addition to the utilitarian and egalitarian cases, we can use $p \in [2..\infty)$ for computing different consensus. To illustrate the semantics of the ethical principle p and its impact on the consensus, we show a test case (with fabricated data) in Fig. 1, which plots the judgements of 25 individuals on two objects. The circles represent the individuals' judgements $R_i[1]$ and $R_i[2]$ about objects 1 and 2 within the x and y axis, respectively.

We clearly distinguish two groups of individuals: (1) a clustered set of individuals that represent the majority with values for x and y smaller than 4, and (2) a few individuals representing outliers distant from the majority (each shown in a crossed circle). In addition, the squares represent the position of the computed judgement consensus with different ethical principles, p's, whose values are represented with a colour scale from blue ($p = 1$) to red ($p = \infty$). As we can observe, the utilitarian consensus ($p = 1$) is at the centre of the majority. As we increase p, the consensus moves towards the outliers, converging to the egalitarian solution ($p = \infty$), which reduces the distance of the consensus to the worst-off member of the society.

In Section 4, we employ the general distance function in Eq. (6) to pose our problem of aggregating value systems as that of computing a consensus. By leveraging this general distance function, Section 6 analyses how the ethical principle determined by p (including the utilitarian and egalitarian principles) affects the consensus in particular case studies.

3. Value systems: From theory to practice

Moral values are the principles deemed valuable by society [68] and involved in ethical choices [69]. Our preference over relevant, competing values guides our decision-making process [31]. To this extent, *value systems* are the structures employed to represent our moral values

and associated preferences [29–31]. Rankings are typically used to describe value preferences because they are the less restrictive preference structure satisfying totality. However, recent works [8,38] contend that representing value systems solely as rankings is incomplete for concrete applications.

Computer scientists [8,34] and social scientists [32,33,35] argue that value systems are *contextual*. That is, the way in which we reason about and prioritise values is influenced by the context we are in. For instance, one may value freedom over safety but prioritise safety over freedom during a global pandemic [8]. Further, thinking about values is challenging for humans since it involves significant cognitive effort [70,71]. Thus, allowing humans to express their preferences over value-laden contextual alternatives (as opposed to competing values) is both easier and more insightful [72].

Along the lines of Gabriel [6], both Liscio et al. [8] and Lera-Leri et al. [38] consider value systems as *pluralistic*. As argued by Gabriel [6], humans hold various reasonable but contrasting beliefs about values. Therefore, we live in a pluralistic world where individuals ascribe to different value systems. Thus, individuals from different cultural backgrounds might judge the same action differently regarding the same moral value, which Lera-Leri et al. [38] describe as having different *value interpretations*. For instance, if we consider the moral value of *respect* in the context of a funeral, Western cultures consider wearing black as promoting the value (and failing to do it as demoting it), whereas Asian cultures favour a white dress code.

Considering the *contextuality* and *pluralism* of value systems, we identify four key elements that a value system must include:

1. a set of **values** relevant to the context under discussion;
2. a set of contextual **alternatives** (e.g., actions, policies) over which an individual ought to take value-laden choices;
3. a **relationship between alternatives and values** that captures value interpretation by specifying the degree to which an individual deems an alternative as *promoting* or *demoting* a value; and
4. a set of individual **preferences over values**.

On the one hand, the presence of alternatives and the relationship between alternatives and values delimit a context for situated value reasoning. On the other hand, the individual interpretation of the relationship between alternatives and values and consequent individual value preferences reflect the pluralism of values. To this extent, Chisholm [73] links these two aspects (i.e., alternatives or actions and values) by arguing that individuals may judge an action as either good or bad to perform (or not to perform) depending on the value under consideration.

Our goal is to aggregate value systems to obtain a consensus value system. To this end, we introduce two case studies aimed at estimating value systems (Sections 3.1 and 3.2). In each case study, we formally define the corresponding value system and show how individual value systems are obtained. Subsequently, in Section 6, we compute the consensus value system by aggregating both value interpretations and value preferences.

3.1. Participatory value evaluation

Participatory Value Evaluation (PVE) [40] is a digital framework for eliciting citizens' preferences over governmental policy options. We use data from a PVE conducted in 2020 to gauge the opinion of the residents of a municipality in the Netherlands over energy policies [41].

The main question asked in this PVE was: *what do you find important for future decisions on energy policy?* As potential answers, six *policy options* were developed by policymakers in collaboration with a panel of 45 citizens. Each PVE participant was asked to distribute 100 points among the six options and then to motivate each option to which they had assigned points with a textual *motivation*. Table 1 shows the six policy options and the mean points allocated to each option by the PVE participants.

The mean points allocated to each option in a PVE indicate *what* options the PVE participants prefer. However, these points do not indicate *why* the participants have that preference. To answer the why question, Siebert et al. [43] estimate the value system underlying each participant's preferences.

In Section 3.1.1, we describe how the previously introduced value system's four key components connect to this case study and provide a formal definition of a value system. In Section 3.1.2, we introduce the method that Siebert et al. [43] propose for estimating an individual's value system from their PVE data.

3.1.1. Definition of value system

We adapt the definition of value system that Siebert et al. [43] propose to incorporate the four key components introduced earlier.

1. The *value list* V is the set of values relevant to the discussion.
2. The *alternatives* over which participants make value-laden choices are the policy options presented in the PVE (Table 1): $O = \{o_1, \dots, o_6\}$.
3. The *relationship between alternatives and values* captures value interpretation by means of a binary individual value-option matrix VO_i with $|V|$ (number of values) rows and $|O|$ (number of options) columns, where:

$$VO_i(v, o) = \begin{cases} 1, & \text{if value } v \text{ is relevant for an individual } i \text{ for option } o \\ 0, & \text{otherwise.} \end{cases}$$

   In Section 3.1.2, we explain how an individual's VO_i is calculated based on their answers to the survey.
4. The *value preferences* are an individual's ranking R_i of V, which is a reflexive, transitive, and total binary relation, noted as $v_a \succeq v_b$. Given $v_a, v_b \in V$, if $v_a \succeq v_b$, we say v_a is more preferred than v_b. If $v_a \succeq v_b$ and $v_b \succeq v_a$, then we note it as $v_a \sim v_b$ and consider v_a and v_b indifferently preferred. However, if $v_a \succeq v_b$ but it is not true that $v_b \succeq v_a$ (i.e., $v_a \neq v_b$), then we note it as $v_a \succ v_b$.

Considering the four aspects, we define a value system in a PVE as follows.

Definition 1. A value system in a Participatory Value Evaluation is a tuple $\mathcal{V}_i = \langle V, O, VO_i, R_i \rangle$, where V is a non-empty set of values, O is a set of available policy options, VO_i is a matrix describing the relevance of a value for an individual for a given option, and R_i is the ranking of V which represents an individual's value preference.

3.1.2. Value system estimation

The designers of the energy policy PVE identify the relevant values by analysing the participants' motivations for policy choices using a grounded theory approach. The five most commonly mentioned values they identified are: cost-effectiveness (v_1), nature and landscape preservation (v_2), leadership (v_3), cooperation (v_4), and self-determination (v_5). However, these five values may not be relevant to each of the six policy options in Table 1.

Siebert et al. [43] propose computational methods for estimating a participant's value system based on the choices and motivations they provide in the PVE. They compute an initial VO matrix, shown in Table 2, as the first guess of value preferences based on the available choices in the PVE. It is intended to be the starting point for estimating the individual participants' value-option matrices (VO_i) and value rankings (R_i).

Siebert et al. [43] propose four methods for estimating participants' value systems. Their approach is based on the theory that "valuing is deliberatively consequential" [74]. That is, if a participant's choice is based on a deliberation of value preferences, the value preferences can be recognised in the motivation provided for the choice. The best-performing approach is the sequential combination of three proposed methods (MO, MC, and TB):

Table 1
Policy options in the Súdwest-Fryslân PVE.

Policy option	Description	Mean points
o_1	The municipality takes the lead and unburdens you	29.05
o_2	Inhabitants do it themselves	21.72
o_3	The market determines what is coming	9.39
o_4	Large-scale energy generation will occur in a small number of places	15.01
o_5	Betting on storage (Súdwest-Fryslân becomes the battery of the Netherlands)	12.96
o_6	Become a major energy supplier in the Netherlands	4.71

Table 2
VO matrices representing the relationship between alternatives and values.

(a) Initial VO for the energy transition PVE

Values \ Options	o_1	o_2	o_3	o_4	o_5	o_6
v_1	1	1	1	1	1	1
v_2	1	1	0	1	1	1
v_3	1	1	1	0	0	0
v_4	1	1	1	1	0	1
v_5	1	1	0	0	1	0

(b) Example of estimated individual VO_i

Values \ Options	o_1	o_2	o_3	o_4	o_5	o_6
v_1	1	1	1	1	0	1
v_2	1	1	0	1	1	0
v_3	1	1	1	0	0	0
v_4	1	1	1	0	1	1
v_5	1	1	0	0	1	0

- MO addresses inconsistencies between motivations provided for different choices. For example, consider that an individual selected options o_4 and o_6, for which v_2 and v_4 are deemed relevant. Further, assume that they motivated o_4 with value v_2 and o_6 with value v_4. Following the notion of valuing as deliberatively consequential, from the two motivations, one can both infer that $v_2 \succ v_4$ and that $v_4 \succ v_2$. Thus, the method MO updates VO_i by considering irrelevant the value that is part of the inconsistency but was not mentioned in the motivation (in practice, by setting the cells (v_4, o_4) and (v_2, o_6) to 0, as in Table 2).
- MC targets inconsistencies between choices and motivations. Assume that a participant allocates some points to option o_5 where, according to VO, v_1 is relevant, but v_4 is not, but mentions v_4 in the motivation. Then, MC adjusts VO_i to prefer v_4 over v_1 for o_5 (in practice, by setting the cell (v_1, o_5) to 0 and the cell (v_4, o_5) to 1, as in Table 2).
- TB reduces the number of indifferent preferences in an individual's value ranking. First, it computes the importance of values for an individual by weighing the values supported in VO_i with the points that the participant assigned to the options. For instance, consider that individual i distributes their points to the six policy options as: $\{30, 40, 10, 20, 0, 0\}$. The multiplication of VO_i by the vector of this policy scores results in a vector of importance scores for values in V: $\{100, 90, 80, 80, 70\}$. Then, the individual's value ranking R_i is inferred by ordering the values in V according to their importance score: $v_1 \succ v_2 \succ v_3 \sim v_4 \succ v_5$, where $v_a \sim v_b$ indicates that there is a tie, i.e., the participant has no preference between v_a and v_b. However, if one of the motivations provided by the participant mentions v_4 but no motivations mention v_3, then the TB method breaks the tie by setting $v_4 \succ v_3$, thus resulting in $R_i : v_1 \succ v_2 \succ v_4 \succ v_3 \succ v_5$.

3.2. European value study

The European Values Study (EVS) [42] is a large-scale survey research programme on European values. It collaborates with the World Values Survey [75]. The programme provides data about the variety of positions that citizens from different European countries have on basic values such as well-being, solidarity, and democracy. Although the EVS survey covers a wide range of questions and values, here we focus on two values and three questions for 34 European countries. Our goal is not to comprehensively analyse European values but to illustrate our aggregation approach with a simplified example.

3.2.1. Definition of value systems

We adopt the definition of Value System $\mathcal{V}$ for the EVS from Lera-Leri et al. [38] and identify the four key elements of this value system.

1. $V = \{v_1, \ldots, v_n\}$ is the set of values relevant to European citizens.
2. The alternatives are a set of actions, $A = \{a, \ldots, a_m\}$, which give information about the value interpretation of citizens.
3. We relate values with actions to interpret each value $v \in V$ for every country i. To do so, for each value $v \in V$ and the actions in A, we define the action judgement function $\alpha^i_v : A \rightarrow [-1, 1]$ for country i as the function that evaluates the promotion or demotion of value v when performing action $a \in A$. These evaluations are real numbers in the interval $[-1, 1]$: a positive number indicates the degree to which a value is being promoted, whereas a negative one indicates demotion. For instance, in the "funeral dress code" example from Section 3, a western-raised individual would consider wearing black clothes (wbc) as an action that promotes the value respect ($\alpha^i_{respect}(wbc) > 0$), and wearing colourful clothes (wcc) is an action that demotes the value respect ($\alpha^i_{respect}(wcc) < 0$). We represent such evaluations in a value-action matrix such that
$$VA_i = \begin{bmatrix} \alpha^i_{v_1}(a_1) & \cdots & \alpha^i_{v_1}(a_m) \\ \vdots & \ddots & \vdots \\ \alpha^i_{v_n}(a_1) & \cdots & \alpha^i_{v_n}(a_m) \end{bmatrix}, \quad (5)$$
where each row corresponds to the action judgement functions of a value.
4. We define a country i's value preferences via a preference matrix $P_i \in [0, 1]^{n \times n}$, where $P_i[v_j, v_k] \in [0, 1]$ represents i's graded preference when comparing v_j and v_k. The value preferences in P_i are *pairwise comparisons*, where 0.5 stands for indifference. For instance, individuals from a country may prefer environmental protection over economic development with a grade of 0.75.

Given these elements, we formally define the value system as follows.

Definition 2. Given a set of values V and a set of actions A, a value system for the European Values Study is a tuple $\mathcal{V}_i = \langle V, A, VA_i, P_i \rangle$ for individual i, where VA_i is the value-action matrix containing the evaluation of values with relation to actions, and P_i is the preference matrix containing the preference pairwise comparisons between values.

Table 3
Value preferences (columns 2 and 3) and value judgement functions (columns 4–7) for a subset of countries in the EVS.

Country	$P_i[rl, pr]$	$P_i[pr, rl]$	$\alpha^i_{rl}(ho)$	$\alpha^i_{rl}(dv)$	$\alpha^i_{pr}(ho)$	$\alpha^i_{pr}(dv)$
CH	0.32	0.68	0.01	0.20	0.30	0.56
CZ	0.22	0.78	−0.09	0.01	0.12	0.32
DE	0.36	0.64	0.20	0.25	0.42	0.59
DK	0.24	0.76	0.25	0.53	0.46	0.75
EE	0.26	0.74	−0.52	−0.07	−0.23	0.24
ES	0.40	0.60	0.15	0.20	0.44	0.50
FI	0.35	0.65	0.02	0.30	0.32	0.62
FR	0.36	0.64	0.29	0.26	0.51	0.54
GB	0.38	0.62	0.20	0.20	0.39	0.50
GE	0.94	0.06	−0.71	−0.50	−0.60	−0.32
HR	0.61	0.39	−0.41	−0.17	−0.02	0.28
NL	0.32	0.68	0.26	0.19	0.52	0.63

3.2.2. Value system estimation

We resort to the EVS data [76] to create the value system of each country. We consider two values: *religiosity* (rl) and *permissiveness* (pr).[3] We characterise these two values in terms of their action judgement functions. For simplicity, we consider the judgement of two actions: *homosexual couples' parenthood* (ho) and *divorcing* (dv). To characterise Europeans' position on religiosity, we consider the EVS question "Q1F: How important is religion in your life?" and partition possible answers so that we can discern the percentage of citizens who consider religion important from the ones who do not. Columns 2 and 3 of Table 3 show the respective percentages per country, which we also interpret as the degree of preference of each value. For conciseness, Table 3 lists 12 out of the 34 countries considered in our computation. Formally, we denote the preference degree of value rl over pr in country i as $P_i[rl, pr]$. Conversely, we denote as $P_i[pr, rl]$ the preference degree of pr over rl. As a consequence, we assume that those countries in which religion is important for the majority of the population (i.e., $P_i[rl, pr] > P_i[pr, rl]$) will prefer religiosity over permissiveness, whereas we consider that permissiveness is preferred over religiosity if $P_i[rl, pr] < P_i[pr, rl]$.

Next, we employ two additional EVS questions to characterise the value judgement functions of the values under consideration: "Q27A: How much do you agree or disagree with the statement: Homosexual couples are as good parents as other couples?" and "Q44G: Can divorce be always justified, never justified, or something in between?". By correlating these answers with those about religion, we obtain the judgements of religious citizens of each country on homosexual parenthood and divorce (columns 4–5 of Table 3). Similarly, we obtain the judgements of non-religious people (columns 6–7 of Table 3).

4. Formalising the aggregation of value systems

After formally describing a value system $\mathcal{V}$ and two real-world examples, we proceed to the problem of aggregating value systems, $\mathcal{V}_1, \dots, \mathcal{V}_N$, of individuals $i = 1, \dots, N$. Specifically, we compute a value system $\mathcal{V}_S$ that *best* represents the aggregation of $\mathcal{V}_1, \dots, \mathcal{V}_N$ according to an *ethical principle* p.[4]

Recall that a value system has four key components. We assume that the first two components, the set of values and the set of alternatives, are the same across all individuals. So it is not necessary to aggregate them. Then, the problem of aggregating multiple value systems boils down to aggregating the last two components, the relationship between values and alternatives and the value preferences, which are specific to the individuals.

We pose the overall aggregation problem as a two-step procedure to compute: (1) a *consensus* values-alternatives relation, and (2) a *consensus* of preferences over values. We represent the objects for aggregation (i.e., the individual values-alternatives relations and the individual preferences over values) as q-dimensional vectors. This representation provides sufficient expressiveness since it generalises 1-dimensional vectors and matrices, which are commonly used in the literature [44, 55–57] and in the test cases described in Sections 3.1 and 3.2 for these purposes.

Considering a unique general representation for individual objects allows us to pose the aggregation problem in a unique general way. Following the social choice literature (Section 2), we cast this problem as the minimisation of a distance measure defined along the lines of Eqs. (1) and (2) (for 1 and 2-dimensional cases, respectively). To accommodate a q-dimensional representation, we generalise these distance functions as:

$$U_p = \left[\sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \cdots \sum_{j_q=1}^{J_q} w_i |T_i[j_1, \dots, j_q] - T_S[j_1, \dots, j_q]|^p \right]^{1/p}, \tag{6}$$

where $T_i[j_1, \dots, j_q]$ is the judgement or preference value provided by the ith member of the society to the particular combination over q features.[5] Notice that the generalisation that we propose in Eq. (6) does not assume any underlying vectorial structure. We propose such generalisation to aggregate vectors, matrices, $q-$dimensional vectors, and even scalars when considered as one-component vectors.

Following Definition 1, Eq. (6) can be readily particularised for the PVE case study to define the distance between the individual value-option matrices VO_i and the aggregated value-option matrix VO_S as

$$U_p^{(VO)} = \left[\sum_{i=1}^{N} w_i \sum_{j=1}^{|V|} \sum_{k=1}^{|O|} \left| VO_i[v, o] - VO_S[v, o] \right|^p \right]^{1/p}. \tag{7}$$

[3] Although the values of *religiosity* [77] and *permissiveness* [78] can be related to the values of *tradition* and *tolerance* from the Schwartz's revised model of values [79], we choose them to fit EVS's data better. In fact, one may even think that secularism seems a better alternative to permissiveness when comparing it to religiosity. However, we argue that permissiveness [78] is better suited, as it is specifically related to sexual freedom [80], and the data from EVS we use relates to homosexual couples and divorce.

[4] We refer to p as the ethical principle used to compute the aggregation, in accordance with the social choice literature (Section 2). This should not be confused with the objects of our aggregation, i.e., the value systems.

[5] The judgements or preferences of the individuals $T_i[j_1, \dots, j_q]$ can be represented with binary, integer or real numbers, depending on the case study domain.

From Eq. (6), we can also derive a definition for the distance between the individual rankings R_i and the aggregated ranking R_S as

$$U_p^{(R)} = \left[\sum_{i=1}^{N} w_i \sum_{j=1}^{|V|} \left|R_i[j] - R_S[j]\right|^p\right]^{1/p}. \tag{8}$$

Definitions for the EVS case can be derived similarly. We do not report these definitions for the sake of conciseness.

For a general value system defined in Section 3, we denote the distance function referring to the relationships between values and alternatives, and the value preferences as $U_p^{(3)}$ and $U_p^{(4)}$, respectively. Then, we can pose *value system aggregation* as a two-step problem aiming to compute

$$T_S^{(3)} = \arg\min U_p^{(3)}, \tag{9}$$

$$T_S^{(4)} = \arg\min U_p^{(4)}, \tag{10}$$

where $T_S^{(3)}$ and $T_S^{(4)}$ denote the consensus among values-alternatives relationships and value preferences, respectively.

5. An ℓ_p-regression approach to aggregate value systems

We show how Eqs. (9) and (10) can be cast as ℓ_p-regression (also known as *norm approximation* [46]) problems. Such a transformation yields obvious computational benefits as it allows us to efficiently solve the above-mentioned optimisation problems for any p, as explained in Section 5.1.

Since solving Eqs. (9) and (10) can be seen as the minimisation of the same general q-dimensional distance function (i.e., Eq. (6)), in Theorem 5.1 we directly consider this problem.

Theorem 5.1. *Computing the solution $T_S = \arg\min U_p^{(T)}$ is equivalent to computing the solution x of the ℓ_p-regression problem*

$$\text{minimise } \|Ax - b\|_p, \tag{11}$$

where $A \in \mathbb{R}^{N \cdot J_1 \cdots J_q \times J_1 \cdots J_q}$ and $b \in \mathbb{R}^{N \cdot J_1 \cdots J_q}$ are

$$A = \begin{bmatrix} w_1^{1/p} \cdot I \\ \vdots \\ w_N^{1/p} \cdot I \end{bmatrix}, \quad b = \begin{bmatrix} w_1^{1/p} \cdot \overrightarrow{T_1} \\ \vdots \\ w_N^{1/p} \cdot \overrightarrow{T_N} \end{bmatrix},$$

$I \in \mathbb{R}^{J_1 \cdot J_2 \cdots J_q \times J_1 \cdot J_2 \cdots J_q}$ is the identity matrix of size $J_1 \cdot J_2 \cdots J_q$, $\overrightarrow{(\cdot)}$ is the vectorisation operation that turns a $q-$dimensional vector into a $1-$dimensional vector, and the p-norm $\|x\|_p$ of a vector x is defined as $\|x\|_p = \left(\sum_i |x[i]|^p\right)^{1/p}$.

Proof. As a first step, we rewrite Eq. (6) as

$$\left[\sum_{i=1}^{N} w_i \sum_{h=1}^{J_1 \cdot J_2 \cdots J_q} \left|T_i[h] - T_S[h]\right|^p\right]^{1/p} \tag{12}$$

and, subsequently, as

$$\left[\sum_{i=1}^{N} \left\|w_i^{1/p} \cdot \overrightarrow{T_i} - w_i^{1/p} \cdot \overrightarrow{T_S}\right\|_p^p\right]^{1/p}. \tag{13}$$

To express Eq. (13) as an ℓ_p-regression problem, we define $A \in \mathbb{R}^{N \cdot J_1 \cdot J_2 \cdots J_q \times J_1 \cdot J_2 \cdots J_q}$ and $b \in \mathbb{R}^{N \cdot J_1 \cdot J_2 \cdots J_q}$ as

$$A = \begin{bmatrix} w_1^{1/p} \cdot I \\ \vdots \\ w_N^{1/p} \cdot I \end{bmatrix}, \quad b = \begin{bmatrix} w_1^{1/p} \cdot \overrightarrow{T_1} \\ \vdots \\ w_N^{1/p} \cdot \overrightarrow{T_N} \end{bmatrix},$$

We can finally formulate the problem of minimising Eq. (6) as

$$\text{minimise } \|Ax - b\|_p.$$

The solution of the above-defined problem (i.e., the vector x) is T_S. □

5.1. Solving the ℓ_p-regression problem

We now discuss a computational solution to Eq. (11). This solution applies to aggregating both case studies' value-alternative relations and value preferences, as particular cases of Eq. (11).

For $p = 1$, Eq. (11) represents an *absolute residuals approximation* problem. For $p = \infty$, we are dealing with a *Chebyshev approximation* (or *Min-Max approximation*) problem. In both cases, Eq. (11) can be solved via *Linear Programming* [46]. For $p = 2$, Eq. (11) can be solved analytically by treating it as *Least Squares* problem,[6] whose optimal solution is

$$x = (A^T A)^{-1} A^T b. \tag{14}$$

We employ this analytical solution in Theorem 5.2, where we show that, for $p = 2$, the aggregation of a general q-dimensional vector can be obtained as the *weighted arithmetic mean* of the individual q-dimensional vectors $(T_1, \ldots, T_N)$.

Theorem 5.2. *For $p = 2$, T_S can be analytically computed as the weighted arithmetic mean of the individual T q-dimensional vectors $(T_1, \ldots, T_N)$, where the weights are $w_1, \ldots, w_N$.*

Proof. As a first step, we explicitly compute $(A^T A)^{-1}$ as

$$\begin{aligned}(A^T A)^{-1} &= \left(\begin{bmatrix} w_1^{1/2} \cdot I & \cdots & w_N^{1/2} \cdot I \end{bmatrix} \begin{bmatrix} w_1^{1/2} \cdot I \\ \vdots \\ w_N^{1/2} \cdot I \end{bmatrix}\right)^{-1} \\ &= \left(\sum_{i=1}^{N} w_i \cdot I\right)^{-1} = \left(\sum_{i=1}^{N} w_i\right)^{-1} I.\end{aligned} \tag{15}$$

Notice that Eq. (15) is a diagonal matrix whose elements in the diagonal are all equal to the inverse of the sum of the weights. By making use of the above result, we explicitly compute x as

$$\begin{aligned}x &= \left(\sum_{i=1}^{N} w_i\right)^{-1} \overbrace{\begin{bmatrix} w_1^{1/2} \cdot I & \cdots & w_N^{1/2} \cdot I \end{bmatrix}}^{A^T} \begin{bmatrix} w_1^{1/2} \cdot \overrightarrow{T_1} \\ \vdots \\ w_N^{1/2} \cdot \overrightarrow{T_N} \end{bmatrix} \\ &= \frac{\sum_{i=1}^{N} w_i \cdot \left[\overrightarrow{T_i}\right]}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} w_i \cdot \overrightarrow{T_i}}{\sum_{i=1}^{N} w_i}.\end{aligned} \tag{16}$$

Thus, each element of x (i.e., of T_S) is the weighted mean of the corresponding elements of T_i, according to the weights $w_1, \ldots, w_N$. □

For any $p \notin \{1, 2, \infty\}$, Eq. (6) represents a non-linear problem. Nonetheless, by exploiting the structure of Eq. (11) as an ℓ_p-regression problem, we can solve it for any p. To do so, we choose the state-of-the-art *Iteratively Reweighted Least Squares* (IRLS) algorithm [45], the only approach for ℓ_p-regression that is guaranteed to converge for any value of p.[7]

6. Experimental results

We empirically investigate our method for computing the consensus value systems for the case studies in Sections 3.1 and 3.2. Our analysis shows that our approach is flexible enough to aggregate value preferences encoded differently (i.e., as value rankings in the PVE and pairwise comparisons in the EVS case study). In particular, we aim to:

[6] The Least Squares problem is obtained by squaring the objective of the original ℓ_2-regression problem. The obtained problem is equivalent to the original one (i.e., it has the same optimal solution), but it has the advantage that it can be solved analytically by expressing the objective as a convex quadratic function [46].

[7] Our source code: https://github.com/RogerXLera/ValueSystemsAggregation. We use the publicly-available IRLS code [45]: https://github.com/fast-algos/pIRLS.

1. Illustrate how the choice of a given ethical principle (from 1 to ∞) affects the resulting consensus value system;
2. Understand the impact of employing different social weights for the individual value systems on the resulting consensus value system;
3. Analyse the classification of the available ethical principles based on the *proximity* of the consensus they produce with respect to the utilitarian and egalitarian consensuses; and
4. Study at a fine-grained level how the relationship between the distribution of individual value systems and the resulting consensus varies (with respect to the majority and the minority) as the value of the ethical principle increases.

Finally, we distill our empirical analysis to delineate the guidelines a policymaker should follow in choosing an ethical principle when performing the aggregation of ethical principles.

In Sections 6.1 and 6.2, we analyse the results for the two case studies, respectively, to address the first two goals. Then, Section 6.3 tackles the characterisation of the space of ethical principles as pursued by the third goal. Finally, Section 6.4 investigates the relationship between consensus and individuals as the ethical principle changes (i.e., 4th goal).

6.1. Case study: Participatory value evaluation

The PVE case study comprises five values, six options, and 795 individual value systems estimated via the method in Section 3.1.2. In aggregating these value systems, we treat all individuals as equal within society. Thus, all individuals have the same weight in the distance functions employed to compute the consensus among values-alternatives relationships and value preferences in Eqs. (9) and (10), i.e., $\forall i \in \{1, 2, \dots, 795\}, w_i = 1$.

6.1.1. Analysing the consensus on value preferences

We show the results of aggregating the rankings over values corresponding to the 795 individual value systems according to different ethical principles.

Encoding individual value preferences. First, we build the ranking of values for each individual. For each individual i, we set $R_i[v_j]$ to the position of value v_j in i's ranking. For two values, v_j and v_k, that are equally preferred, we assign the same position within the ranking, i.e., $R_i[v_j] = R_i[v_k]$.

Aggregating individual value preferences. Next, we compute the consensus ranking R_S by solving Eq. (10) for distance function in Eq. (8), following the method in Section 5.1. When computing the consensus, we do not enforce the position of a value j to be an integer. Thus, for instance, a consensus of $R_S[v_j] = 2.5$ indicates that the society considers that value v_j has a position between the second and third position in the consensus ranking. We consider that value v_j is preferred over value v_k ($v_j \succ v_k$) when their positions in the consensus ranking differ such that $R_S[v_k] - R_S[v_j] > \epsilon$, where ϵ is a small positive number ($\epsilon = 0.05$ in our experiments). Otherwise, we say the two values are equally preferred or indifferent ($v_j \sim v_k$).

Table 4 shows the consensus rankings resulting from the aggregation of individual rankings for different ethical principles (p): from 1 (utilitarian) to 10, and ∞ (egalitarian). Each column $R_S[v_i]$ indicates the position of value v_i in the ranking computed by our aggregation. Note that we obtained a partial ranking as a consensus ranking for each ethical principle. That is, the order (preferences) between values in each consensus ranking is not strict since it contains ties between values. For instance, in the first row, value v_1 is equally preferred to value v_2 (because $R_S[v_1] = R_S[v_2] = 2$).

From Table 4, we distinguish three types of consensus rankings.

- $p = 1$ (utilitarian): v_1 and v_2 are equally preferred. They are both preferred over the others (v_3, v_4, v_5), which are equally preferred.
- $p \in [2..10]$ (intermediate): v_1 becomes more preferred than v_2, while both are still more preferred than the rest of values (v_3, v_4, v_5). The indifference between v_3, v_4 and v_5 holds.
- $p = \infty$ (egalitarian): all values are equally preferred.

We make two interesting observations from Table 4. First, recall (Section 5.1) that for $p = 2$, the consensus ranking results from computing the mean of the individual rankings to aggregate. Thus, our consensus ranking for $p = 2$ is the same as obtained by Siebert et al. [43], which solely employs the mean of individual rankings to compute a consensus ranking. Second, the consensus position R_S for *all* moral values converges to 3 (central position in the ranking) as the value of parameter p increases.

6.1.2. Analysing the consensus on values-alternatives relationships

After analysing the impact of choosing different ethical principles on the consensus value preferences, we now discuss the results of the consensus value-option matrices. Recall from Section 3.1.2 that each individual holds a different view of the relationship between values and options. Computing the consensus on the value-option matrices unveils the most representative value-option matrix (i.e., the most representative interpretation of values) on which the consensus value preferences apply. Recall (Section 3.1.1) that the binary value-option matrix VO_i indicates whether a value v is promoted by option o or not according to individual i. We remark that considering binary matrices for expressing the relationship between values and alternatives is the result of applying the methodology proposed by Siebert et al. [43]. However, in the EVS case study presented in the following section, we will discuss a different instance of the relationship between alternatives and values that contain more gradual evaluations [−1,1], representing the degree of promotion or demotion of a given value when performing a specific action.

For brevity, we only analyse the consensus in three cases of interest: the utilitarian ($p = 1$), the *mean principle* ($p = 2$), and the egalitarian ($p = \infty$) cases. For ease of understanding, we first discuss the mean principle case.

Mean principle ($p = 2$). Eq. (17) shows the resulting consensus value-option matrix. From Theorem 5.2, we know that aggregating individual value-option matrices results in a value-option matrix containing the mean values of the individual matrices. By multiplying the consensus value-option matrix by 100, we obtain the percentage of individuals considering value v relevant for option o. For example, v_3 is deemed relevant to 45.2% of the participants to justify option o_1; similarly, v_5 and v_2 are related to options o_2 and o_4, respectively, for more than 30% of the participants.

$$VO_S^{(p=2)} = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{array} \begin{array}{c} \begin{array}{cccccc} o_1 & o_2 & o_3 & o_4 & o_5 & o_6 \end{array} \\ \begin{pmatrix} 0.116 & 0.116 & 0.134 & 0.117 & 0.121 & 0.075 \\ 0.065 & 0.039 & 0.015 & 0.349 & 0.037 & 0.061 \\ 0.452 & 0.054 & 0.055 & 0.017 & 0.015 & 0.004 \\ 0.110 & 0.177 & 0.048 & 0.026 & 0.016 & 0.041 \\ 0.048 & 0.396 & 0.012 & 0.011 & 0.028 & 0.022 \end{pmatrix} \end{array} \tag{17}$$

Utilitarian principle ($p = 1$). The consensus value-option matrix VO_S is a zero matrix, i.e., $VO_S[v, o] = 0$ for every value $v \in V$ and option $o \in O$. This indicates that no value $v \in V$ is promoted in any option $o \in O$. Despite Table 2 showing that there are values that are frequently annotated for some options, most individual value-option matrices are sparse (have many zeros). This is reflected in Eq. (17), where there are no options in which a majority of individuals indicated a specific value to be relevant (i.e., no entry of Eq. (17) is larger than 0.5). This leads the method to compute a consensus of zeros because ℓ_1-regression tends to have residuals ($VO_i[v, o] - VO_S[v, o]$) equal to 0. This is well-known in the optimisation literature [46].

Table 4
Computed consensus ranking for different ethical principles p.

p	$R_S[v_1]$	$R_S[v_2]$	$R_S[v_3]$	$R_S[v_4]$	$R_S[v_5]$	Consensus ranking
1	2.00	2.00	3.00	3.00	3.00	$v_1 \sim v_2 \succ v_3 \sim v_4 \sim v_5$
2	2.14	2.53	2.90	2.90	2.91	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
3	2.31	2.65	2.95	2.96	2.94	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
4	2.42	2.73	2.97	2.98	2.96	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
5	2.51	2.78	2.98	2.99	2.97	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
6	2.58	2.81	2.99	3.00	2.98	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
7	2.63	2.84	2.99	3.00	2.98	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
10	2.74	2.89	2.99	3.00	2.99	$v_1 \succ v_2 \succ v_3 \sim v_4 \sim v_5$
∞	3.00	3.00	3.00	3.00	3.00	$v_1 \sim v_2 \sim v_3 \sim v_4 \sim v_5$

Table 5
Value preferences (columns 2, 3, and 8) and value judgement functions (columns 4–7) of the consensus European value system when considering equal social influence.

p	$P_S[rl, pr]$	$P_S[pr, rl]$	$\alpha^S_{rl}(ho)$	$\alpha^S_{rl}(dv)$	$\alpha^S_{pr}(ho)$	$\alpha^S_{pr}(dv)$	Value pref.
1	0.504	0.496	−0.288	0.013	−0.054	0.318	$rl \succ pr$
2	0.535	0.465	−0.224	0.031	−0.032	0.315	$rl \succ pr$
3	0.548	0.452	−0.200	0.029	−0.029	0.280	$rl \succ pr$
∞	0.580	0.420	−0.148	0.027	−0.077	0.159	$rl \succ pr$

Egalitarian principle *($p = \infty$).* The consensus value-option matrix has $VO_S[v, o] = 0.5$ for every value $v \in V$ and option $o \in O$. This is because the egalitarian principle minimises the disagreement with the most displaced elements of the society, i.e., the distance between the consensus for each value-option pair and the individuals assigning the largest and lowest judgement to the value-option pair. Recall from Table 2 that our value-option matrices are binary. By peeking into the individual option matrices, we observe that for every value-option pair (v, o), the maximum $VO_i[v, o]$ is 1 (at least one individual mentions value v to justify option o). The minimum $VO_i[v, o]$ is 0. Hence, $VO_S[v, o] = 0.5$ for every moral value $v \in V$, and option $o \in O$ because 0.5 is the consensus value that minimises the distance to the most displaced individual.

6.2. European value study

We now study the aggregation of value systems for the European Values Study case. Similar to the PVE case, we explore how the resulting value system is affected by the parameters of our aggregation approach. However, unlike the PVE case, in the EVS case, we consider two cases for social influence: (1) where all countries have the same influence ($w_i = 1$), and (2) where social influence w_i of a country i is proportional to the population of that country (i.e., $w_i = \text{population}_i / \sum_j \text{population}_j$).[8]

Recall from Section 3.2 that this case considers 34 European countries, two values (religiosity (rl) and permissiveness (pr)), and two actions (divorce (dv) and parenthood by homosexual couples (ho)). Tables 5 and 6 show the consensus value systems for different ethical principles, considering equal and population-based social influence, respectively. We report results for the utilitarian ($p = 1$) and egalitarian cases ($p = \infty$) and $p = 2$ and $p = 3$. We choose to show these values of parameter p as a change of value preferences can be observed in the consensus computed with $p = 2$ and $p = 3$ in the population-based social influence scenario (Table 6). In both tables, P_S stands for the value of the consensus on moral value preferences, whereas the α values contain the consensus values on moral value-action relationships. For instance, $P_S[rl, pr]$ indicates the consensus grade of preference of the religiosity moral value over the permissiveness moral value. Further, $\alpha^S_{rl}(dv)$ denotes the consensus grade of promotion of the religiosity value when performing action divorce.

[8] The population data is accessed from *Worldometers* (https://www.worldometers.info/world-population/population-by-country) in Sept. 2021.

Equal social influence. In Table 5, all countries have the same social influence. We observe that all the *consensus European value systems* prefer religiosity over permissiveness. This is because the computed preference of religiosity over permissiveness is larger than 0.5 ($P_S[rl, pr] > 0.5$) for all ethical principles (p) considered. However, for the utilitarian case ($p = 1$), this preference is barely noticeable (i.e., both $P_S[rl, pr]$ and $P_S[pr, rl]$ are close to 0.5), whereas the difference between $P_S[rl, pr]$ and $P_S[pr, rl]$ increases as we increase the value of parameter p. This transition happens because, for greater values of p, the consensus tends to reduce the maximum disagreement with the most "extreme" preferences, which in our case is by Georgia ($P_i[rl, pr] = 0.94$). In addition, for all p's the parenthood by homosexual couples is valued negatively (see the values of $\alpha^S_{rl}(ho)$ and $\alpha^S_{pr}(ho)$), whereas divorce is slightly accepted (see the values of $\alpha^S_{rl}(dv)$ and $\alpha^S_{pr}(dv)$).

Population-based social influence. When considering social influence based on population size (Table 6), we observe significant changes compared to the previous case. Specifically, for the utilitarian principle ($p = 1$), permissiveness is preferred over religiosity (since $P_S[pr, rl]$ is larger than $P_S[rl, pr]$) and both adoptions by homosexual couples and divorce promote both values. This consensus shift is due to the social influence that highly populated European countries, such as Germany, France, and Great Britain, wield, which prefer permissiveness over religiosity. Nonetheless, as we increase p, we obtain the same transition towards religiosity as observed in Table 5. Indeed, for values of parameter p equal or larger than 3 we obtain the same trend on all consensus values (for P_S and α) for both the population-based (weighted) and the equal social influence (unweighted) cases. This confirms that the impact of weights (w_i) vanishes when considering high values of parameter p in accordance with Eq. (4).

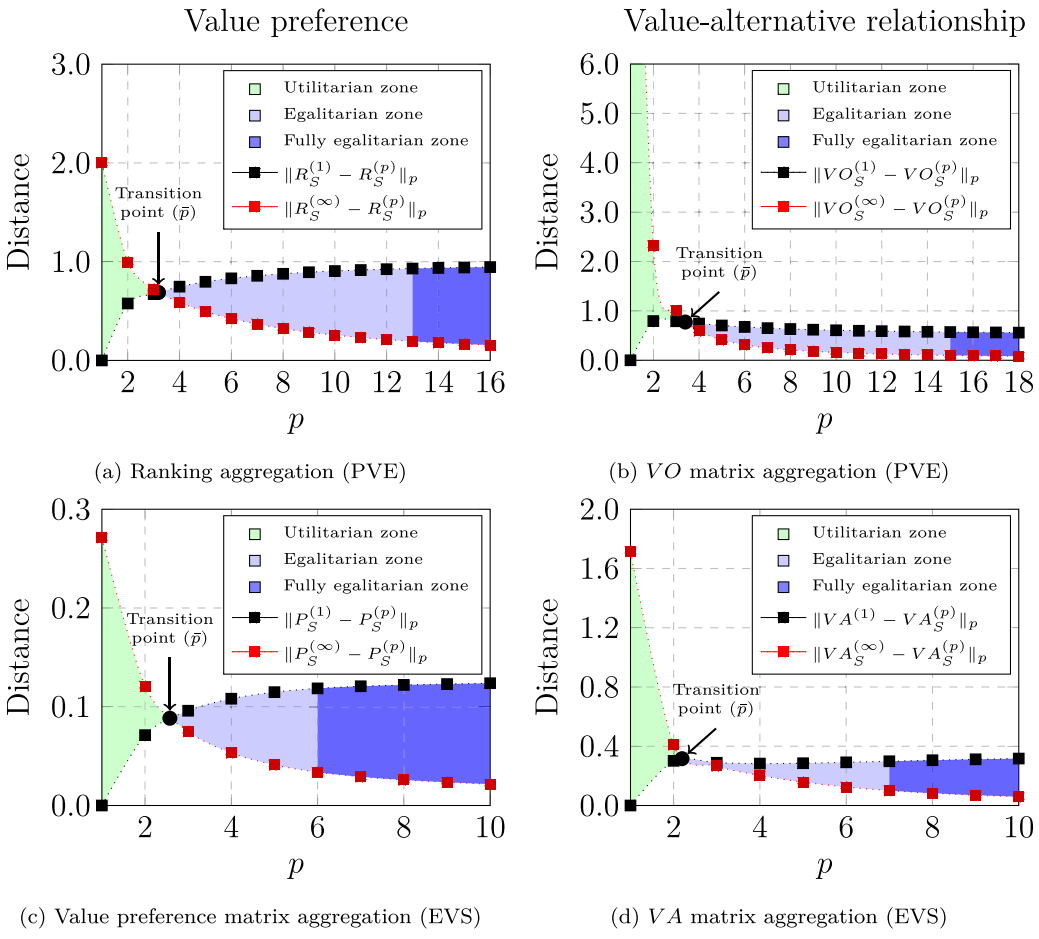
Overall, choosing the aggregation parameters (p and w_i) determines between swaying towards prevailing (majority) value systems or value systems that lay closer to divergent opinions. In the EVS case, we even obtain a swap for the preferences of *religiosity* and *permissiveness*.

6.3. Characterising the space of ethical principles

We characterise the *whole* space of ethical principles (from utilitarian to egalitarian) available to a decision-maker when computing a consensus value system. We do so to determine whether an ethical principle p produces a consensus leaning towards utilitarian ($p = 1$) or egalitarian ($p = \infty$). To achieve our objective in the PVE case, we compute the consensus ranking R_S considering a given p (denoted as $R_S^{(p)}$). We measure the distance between $R_S^{(p)}$ and the one corresponding to $p = 1$ and $p = \infty$, denoted as $R_S^{(1)}$ and $R_S^{(\infty)}$ respectively. Formally,

Table 6
Value preferences (columns 2, 3, and 8) and value judgement functions (columns 4–7) of the consensus European value system when social influence considers population.

p	$P_S[rl, pr]$	$P_S[pr, rl]$	$\alpha^S_{rl}(ho)$	$\alpha^S_{rl}(dv)$	$\alpha^S_{pr}(ho)$	$\alpha^S_{pr}(dv)$	Value pref.
1	0.444	0.556	0.007	0.123	0.187	0.503	$pr \succ rl$
2	0.495	0.505	−0.128	0.111	0.059	0.400	$pr \succ rl$
3	0.521	0.479	−0.154	0.080	0.015	0.364	$rl \succ pr$
∞	0.580	0.420	−0.148	0.027	−0.077	0.159	$rl \succ pr$



(a) Ranking aggregation (PVE)

(b) VO matrix aggregation (PVE)

(c) Value preference matrix aggregation (EVS)

(d) VA matrix aggregation (EVS)

Fig. 2. Distance between the consensus computed according to ethical principle p and the consensuses computed according to $p = 1$ (fully utilitarian, black line) and $p = \infty$ (fully egalitarian, red line). The transition point $\bar{p}$ is the ethical principle producing a consensus equidistant from the fully utilitarian and fully egalitarian ones. Thus, $\bar{p}$ divides the space of ethical principles into an *utilitarian zone* (more similar to the fully utilitarian consensus, green) and an *egalitarian zone* (more similar to the fully egalitarian consensus, light blue). The *fully egalitarian* (dark blue) zone marks the ethical principles that produce a consensus approximately equal (to a small ϵ) to the fully egalitarian one. PVE case study on top, and EVS case study for the population-based influence case at the bottom.

we denote these two distances as $\left\|R_S^{(1)} - R_S^{(p)}\right\|_p$ and $\left\|R_S^{(p)} - R_S^{(\infty)}\right\|_p$. Analogously, we define $\left\|VO_S^{(1)} - VO_S^{(p)}\right\|_p$ and $\left\|VO_S^{(p)} - VO_S^{(\infty)}\right\|_p$ for value-option matrices. In the same vein, for the EVS case, we denote the distances between value preferences as $\left\|P_S^{(1)} - P_S^{(p)}\right\|_p$ and $\left\|P_S^{(p)} - P_S^{(\infty)}\right\|_p$, and denote the distances between value-action matrices as $\left\|VA_S^{(1)} - VA_S^{(p)}\right\|_p$ and $\left\|VA_S^{(p)} - VA_S^{(\infty)}\right\|_p$.

By making use of the above-defined distances, we can determine a *transition point* (denoted as $\bar{p}$) that is the equidistant ethical principle whose computed consensus is between the fully utilitarian and the fully egalitarian consensuses. In addition, we define the *limit point*, $\hat{p}$, as the ethical principle such that all $p > \hat{p}$ produce a consensus that is approximately equal (to a small ϵ) to the fully egalitarian one ($p = \infty$). We compute $\hat{p}$ that satisfies:

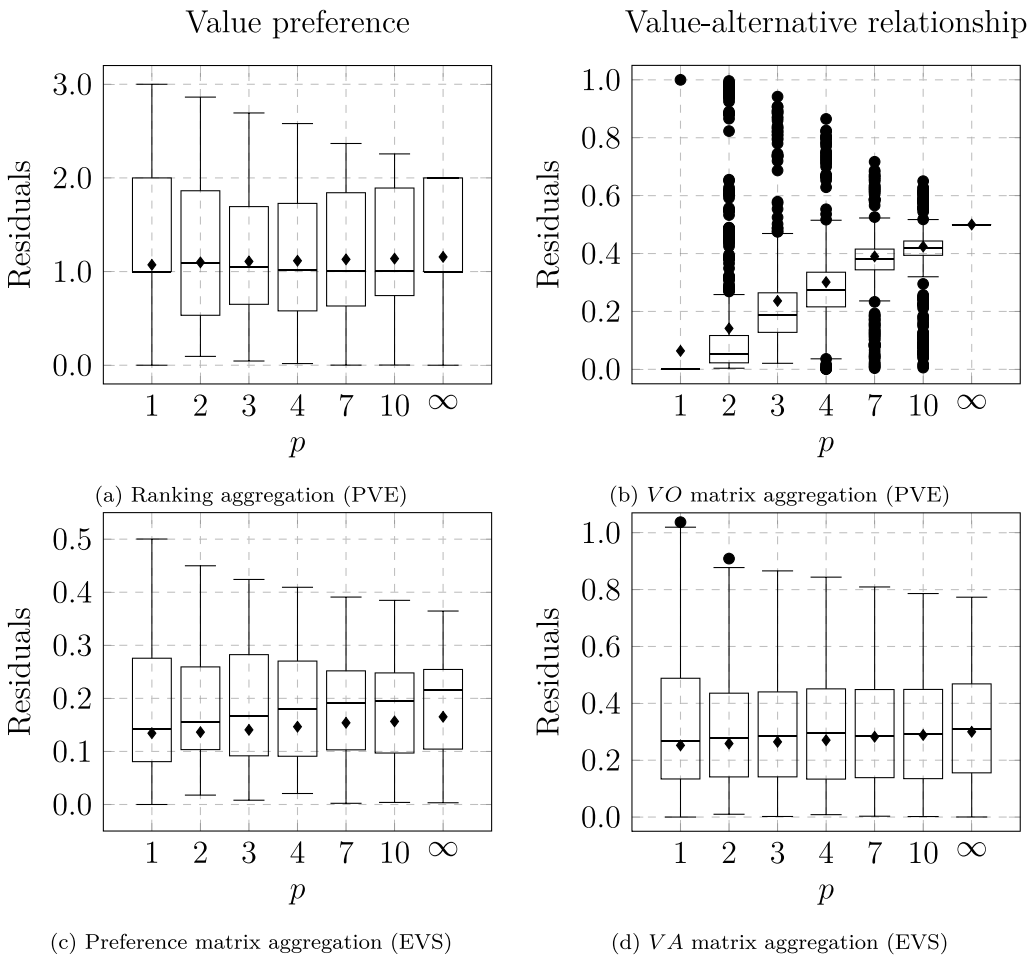
$$\frac{U_{p-1}^{(R)} - U_p^{(R)}}{U_p^{(R)}} < \epsilon,$$

where $U_p^{(R)}$ is the value of the distance function defined by Eq. (8).

Because of the transition and limit points, we can characterise different zones within the space of ethical principles as Fig. 2 illustrates.

- The *utilitarian zone* is composed of all ethical principles leaning towards the fully utilitarian case ($p < \bar{p}$).
- The *egalitarian zone* is composed of all ethical principles leaning towards the fully egalitarian case but before surpassing the *limit point* ($\hat{p}$), i.e., $p \in [\bar{p}..\hat{p}]$.
- The *fully egalitarian zone* is composed of the set of all ethical principles greater than the limit point, i.e., $p > \hat{p}$.

Fig. 2(a) plots the distances between the consensus ranking ($R_S^{(p)}$) with respect to the fully utilitarian ($R_S^{(1)}$) and fully egalitarian consensus rankings ($R_S^{(\infty)}$) as the value of the ethical principle p increases. Fig. 2(b) shows analogous results for aggregating value-option matrices. We observe that the transition point is near 3 ($\bar{p} \sim 3$) for both the ranking and value-option matrix aggregation. However, the limit point



(a) Ranking aggregation (PVE)

(b) VO matrix aggregation (PVE)

(c) Preference matrix aggregation (EVS)

(d) VA matrix aggregation (EVS)

Fig. 3. Boxplots of residuals for different ethical principles in the value preference and value-alternative aggregation for the PVE (top) and EVS (bottom) case study. The diamond represents the mean and the circles represent the outliers.

varies for both cases ($\hat{p} = 13$ for ranking aggregation and $\hat{p} = 15$ for value-option matrix aggregation). This is because such points depend on the data of the problem.

The data dependency is corroborated when considering the EVS case. Figs. 2(c) and 2(d) plot the distance between consensuses for value preference matrix aggregation and values-action matrix aggregation, respectively. In these figures, we observe that the transition and limit points are different with respect to those computed for the PVE case. The transition point for value preference and value-action matrices aggregation is between 2 and 3, whereas the limit point is $\hat{p} = 6$ in Fig. 2(c) and $\hat{p} = 7$ in Fig. 2(d).

Therefore, a given ethical principle p can be interpreted as "more utilitarian" or "more egalitarian" depending on its relative position with respect to the transition point. However, notice that if the transition point $\bar{p}$ turns out to be different for the two aggregations that we perform to compute a consensus value system, then if we choose the same ethical principle p to apply to both aggregations, it may lay in different regions.

6.3.1. Takeaways for decision-makers

The comparison between the two case studies shows that the regions characterising the space of ethical principles depend on the domain and the data. Hence, a decision-maker must carefully examine the different ethical principles zones when choosing the ethical principle p. The decision-maker may desire to choose an ethical principle that leads to either utilitarian or egalitarian consensus for both aggregations (value preferences and value interpretations). However, the same value of parameter p could fall into different zones for moral value preferences and moral value-alternatives relationships aggregation. This motivates the need for characterising the *joint* utilitarian and egalitarian zones for value preferences and value-alternatives relationship aggregations before choosing an ethical principle.

The visual analysis displayed in Fig. 2 intends to provide useful guidance for decision-makers concerned with obtaining a consensus on different value systems following an ethical principle of choice. In general, we propose the following guidelines:

1. Plot the distance between consensuses, as we do in Fig. 2.
2. Plot the utilitarian, egalitarian, and fully egalitarian zones for value preference aggregation and values-alternatives relationship aggregation.
3. Compute the joint utilitarian and egalitarian zones for both value preference and value-alternative relationship. For instance, in the PVE case study (the case for EVS is analogous), we define:
   - $[1, \bar{p}_m)$, where $\bar{p}_m = min(\bar{p}_R, \bar{p}_{VO})$, as the joint utilitarian zone;
   - $(\bar{p}_M, \infty)$, where $\bar{p}_M = max(\bar{p}_R, \bar{p}_{VO})$, as the joint egalitarian zone;
   - $[\bar{p}_m, \bar{p}_M]$ as a *mixed* zone that contains ethical principles that lie in different zones considering the consensus of ranking and value-option matrices.

After following the steps above, we obtain a *joint* space of ethical principles, which is partitioned into three zones of ethical principles:

- Utilitarian: ethical principles for which consensus on value preferences and value-alternatives relationship aggregation are both utilitarian.
- Egalitarian: ethical principles for which consensus on value preferences and value-alternatives relationship are both egalitarian.
- Mixed: ethical principles for which consensus on value preferences and value-alternatives relationship aggregation are not aligned (one is utilitarian, whereas the other is egalitarian).

This joint ethical principle space provides the decision-maker with the necessary information on choosing an ethical principle p to compute the consensus.

6.4. The relationship between consensus and individual value systems

In Sections 6.1 and 6.2, we showed that the choice of the ethical principle impacts the resulting consensus value system. A further *micro-level* analysis helps us investigate how the relationship between the consensus and the individual value systems change as the ethical principle changes. Such a micro-level analysis helps quantify the trade-off when moving away from the majority (utilitarian case) to include the minority (egalitarian case). This analysis complements the one conducted in Section 6.3 (when characterising ethical principles) to help the policymaker choose the ethical principle to employ when aggregating value systems.

In what follows, we analyse the distribution of *residuals*, which represent the *gap* between given consensus value systems and the individual value systems aggregated to obtain the consensus. For instance, when considering the PVE case study, $|R_i[v] - R_S[v]|$ yields the residual for the ranking of individual value system i with respect to the consensus value system regarding value v. Analogously, we can also calculate the residual for the value-option matrix of individual value system i ($|VO_i[v,o] - VO_S[v,o]|$). Fig. 3 plots the distribution of the residuals obtained when computing consensus value systems for both case studies as the value of the ethical principle (p) increases.

First, we focus on the PVE case study. Figs. 3(a) and 3(b) plot the distributions of the residuals for the ranking aggregation and the VO-matrix aggregation, respectively. In Fig. 3(a), we observe that the utilitarian ethical principle ($p = 1$) yields the maximum residuals, hence confirming that the utilitarian principle considers less the individuals within the minority of society. As the value of parameter p increases towards the egalitarian principle, the maximum residuals decrease. For the egalitarian ethical principle ($p = \infty$), we obtain that the maximum residual is 33% smaller than for the utilitarian principle (from 3.0 down to 2.0). This is because the egalitarian ethical principle aims to reduce the distance between the consensus and the most distant individuals of society. Further, we observe that the mean of the residuals gradually increases as the value of parameter p increases (up to an 8% increase for the egalitarian case). The observations above are more pronounced when we analyse the residual distribution for aggregating individual VO matrices using different ethical principles (Fig. 3(b)). On the one hand, the value of the maximum residual halves when moving from $p = 1$ (1.0) to $p = \infty$ (0.5). On the other hand, the mean of the residuals dramatically increases from a value close to 0 until reaching 0.5 for the egalitarian case ($p = \infty$), which amounts to a x7 increase of the mean.[9]

We confirm the observations above when analysing the EVS case study. Figs. 3(c) and 3(d) show the distribution of residuals for the aggregation of individual preference matrices P_i and the aggregation of individual value-action matrices VA_i, respectively for the weighted scenario (different social influences per country). On the one hand, the maximum residual decreases as we increase p for both preference matrix and value-action matrix aggregations. On the other hand, there is a smooth increase of the mean of the residuals as we increase p for both the preference matrix and value-action matrix aggregations. However, the values of residuals differ compared to the PVE case. This is due to the differences in each case study's set of individual value systems.

6.4.1. Takeaways for decision-makers

Our analysis shows that moving away from the majority (utilitarian) to include the minority (egalitarian) leads to an increase in the mean of the residuals and a decrease in the maximum residuals. While the guidelines provided in Section 6.3.1 specify macro-directives for choosing the ethical principle p, the analysis of residuals gives concrete information about the trade-offs between choosing different ethical principles. The decision-maker must weigh the extent to which the increase in the average residual of a large portion of individuals compensates for minimising the distance to the minority (when moving from utilitarian to egalitarian principle).

7. Conclusions and future work

The main contribution of this paper is to provide novel computational means and guidelines for a decision-maker to conduct a principled aggregation of value systems. This is a fundamental step towards designing AI systems that align with a group of individuals with different value systems, namely, towards addressing the *pluralistic value alignment* problem defined by Gabriel [6]. Our contributions make headway to developing trustworthy AI [2] systems that adhere to ethical principles and values.

We show that the problem of computing a consensus value system can be cast as an ℓ_p*-regression* [45] (or *norm approximation* [46]) problem. By doing so, we provide a principled and general theoretical framework to solve the above-mentioned problem for a range of ethical principles—from utilitarian (maximum utility) to egalitarian (maximum fairness)—in a scalable and reliable way, thanks to recent results in the machine learning literature [45]. Importantly, our approach also allows us to compute the consensus for *any* single value of p, which was not possible before for the generator of social choice functions that we employ (González-Pachón and Romero [44] can only deal with $p = 1$ and $p = \infty$).

We study the aggregation of value systems using real-world data from two case studies: the Participatory Value Evaluation (PVE) [41] process and the European Values Study (EVS) [42]. Our empirical evaluation of the case studies draws insights for a decision-maker about how to employ our computational tools to perform value system aggregation. In particular, we show how to proceed to: (i) quantify the impact of choosing different ethical principles on the resulting consensus value system; (ii) characterise the semantics of the available ethical principles (to determine whether a given ethical principle produces a consensus leaning towards the utilitarian or the egalitarian ethical principle); and (iii) quantify the trade-off when moving away from the majority to include the minority. Importantly, our observations vary per case study since they handle different value systems.

In this paper, we treat the aggregation of individual value systems. However, the aggregation must be preceded by the *estimation* of individual value systems [39]. To this end, surveys such as PVE and EVS can be employed to collect value-laden input from the participants, and methods like Siebert et al. [43] (Section 3.1.2) are employed to compound the survey answers into individual value systems. Such methods could be further automated using natural language processing to automatically detect the values that are motivating the survey answers [81,82]. Ultimately, the aggregation of value systems is the result of pipelining the estimation of individual value systems with the computation of a consensus value system.

[9] Recall that value-option matrices are binary. We highlight that we obtain a consensus matrix full of zeros for the utilitarian principle ($p = 1$). Because the value-option matrices are binary, all the residuals are either 0 or 1. In fact, more than the 75% of the residuals are 0. As to the egalitarian principle ($p = \infty$), we obtain a consensus matrix full of 0.5 values. Hence, all the residuals are 0.5.

As future work, we envision three research paths. First, we plan to generalise our framework so that each individual can choose their own ethical principle before aggregating value systems. This might call for developing new social choice functions and computational tools to account for multiple ethical principles. Second, we want to develop a *community-oriented* approach to aggregate value systems. As a first step, we would detect communities of individuals with *similar* value systems. This would allow us to compute the consensus value system per community and, ultimately, the aggregation of the value systems of all the communities. Finally, we plan to explore further connections with the literature on multi-expert decision-making (e.g., [49,50]). Indeed, the distance function employed in our paper can be regarded as a penalty function as defined in [83]. However, profiting from the aggregation functions proposed in that body of work would require investigating their social choice properties. Although this challenge demands additional work beyond this paper's scope, we believe it is worth exploring whether we can benefit from using penalty functions with lower computational costs.

CRediT authorship contribution statement

Roger X. Lera-Leri: Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Enrico Liscio:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Filippo Bistaffa:** Writing – review & editing, Software, Methodology, Conceptualization. **Catholijn M. Jonker:** Writing – review & editing, Funding acquisition. **Maite Lopez-Sanchez:** Writing – review & editing, Writing – original draft, Methodology. **Pradeep K. Murukannaiah:** Writing – review & editing, Methodology. **Juan A. Rodriguez-Aguilar:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis. **Francisco Salas-Molina:** Writing – review & editing, Methodology.

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

Data will be made available on request.

Acknowledgements

The authors were supported by the research projects ACISUD (PID2022-136787NB-I00), VAE (TED2021-131295B-C31), VALAWAI (HE-101070930), Fairtrans (PID2021-124361OB-C33), Yoma OR (OPE02570), TAILOR (952215), and the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research. Maite Lopez-Sanchez belongs to the WAI research group (University of Barcelona), an associated unit to CSIC through the IIIA.

References

[1] R. Chatila, V. Dignum, M. Fisher, F. Giannotti, K. Morik, S. Russell, K. Yeung, Trustworthy AI, in: Reflections on Artificial Intelligence for Humanity, 2021, pp. 13–39.

[2] European Comission, Ethics guidelines for trustworthy AI, 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Online.

[3] European Comission, Artificial intelligence act, 2021, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Online.

[4] Institute of Electrical and Electronics Engineers, IEEE global initiative on ethics of autonomous and intelligent systems, 2019, https://standards.ieee.org/industry-connections/ec/autonomous-systems.html. Online.

[5] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Penguin, 2019.

[6] I. Gabriel, Artificial intelligence, values, and alignment, Minds Mach. 30 (3) (2020) 411–437.

[7] C. Sierra, N. Osman, P. Noriega, J. Sabater-Mir, A. Perello-Moragues, Value alignment: A formal approach, in: Proceedings of Responsible Artificial Intelligence Agents Workshop, 2019.

[8] E. Liscio, M. van der Meer, L.C. Siebert, C.M. Jonker, P.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) (2022) 23, http://dx.doi.org/10.1007/s10458-022-09550-0.

[9] T. Winkler, S. Spiekermann, Twenty years of value sensitive design: A review of methodological practices in VSD projects, Ethics Inf. Technol. 23 (1) (2021) 17–21.

[10] B. Friedman, D.G. Hendry, Value Sensitive Design: Shaping Technology with Moral Imagination, 2019.

[11] P. Noriega, H. Verhagen, J. Padget, M. d'Inverno, Ethical online AI systems through conscientious design, IEEE Internet Comput. 25 (6) (2021) 58–64.

[12] M.O. Riedl, B. Harrison, Using stories to teach human values to artificial agents, in: Proceedings of AAAI Workshop: AI, Ethics, and Society, 2016.

[13] D. Abel, J. MacGlashan, M.L. Littman, Reinforcement learning as a framework for ethical decision making, in: AAAI Workshop: AI, Ethics, and Society, 2016.

[14] Y.-H. Wu, S.-D. Lin, A low-cost ethics shaping approach for designing reinforcement learning agents, in: Proceedings of AAAI Conference on Artificial Intelligence, 2018, pp. 1687–1694.

[15] R. Noothigattu, D. Bouneffouf, N. Mattei, R. Chandra, P. Madan, R. Kush, M. Campbell, M. Singh, F. Rossi, Teaching AI agents ethical values using reinforcement learning and policy orchestration, IBM J. Res. Dev. 63 (4/5) (2019) 6377–6381.

[16] A. Balakrishnan, D. Bouneffouf, N. Mattei, F. Rossi, Incorporating behavioral constraints in online AI systems, in: Proceedings of AAAI Conference on Artificial Intelligence, 2019, pp. 3–11.

[17] M. Rodriguez-Soto, M. Lopez-Sanchez, J.A. Rodriguez Aguilar, Multi-objective reinforcement learning for designing ethical environments, in: Proceedings of International Joint Conference on Artificial Intelligence, 2021, pp. 545–551.

[18] M. Rodriguez-Soto, M. Serramia, M. Lopez-Sanchez, J.A. Rodriguez-Aguilar, Instilling moral value alignment by means of multi-objective reinforcement learning, Ethics Inf. Technol. 24 (1) (2022) 1–17.

[19] N. Ajmeri, Engineering Multi-Agent Systems for Ethics and Privacy-Aware Social Computing (Ph.D. thesis), North Carolina State University, 2018.

[20] J. Szabo, J.M. Such, N. Criado, S. Modgil, Integrating quantitative and qualitative reasoning for value alignment, in: European Conference on Multi-Agent Systems, Springer, 2022, pp. 383–402.

[21] J. Szabo, J.M. Such, N. Criado, Understanding the role of values and norms in practical reasoning, in: Multi-Agent Systems and Agreement Technologies, Springer, 2020, pp. 431–439.

[22] M. Serramia, M. Lopez-Sanchez, J.A. Rodriguez-Aguilar, A qualitative approach to composing value-aligned norm systems, in: Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems, 2020, pp. 1233–1241.

[23] N. Montes, C. Sierra, Value-guided synthesis of parametric normative systems, in: Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems, 2021, pp. 907–915.

[24] N. Montes, C. Sierra, Synthesis and properties of optimally value-aligned normative systems, J. Artificial Intelligence Res. 74 (2022) 1739–1774.

[25] K. Pigmans, N. Doorn, H. Aldewereld, V. Dignum, Decision-making in water governance: From conflicting interests to shared values, in: Responsible Innovation, Springer, 2017, pp. 165–178.

[26] K. Pigmans, H. Aldewereld, V. Dignum, N. Doorn, The role of value deliberation to improve stakeholder participation in issues of water governance, Water Resourc. Manag. 33 (12) (2019) 4067–4085.

[27] C. Flathmann, B.G. Schelble, R. Zhang, N.J. McNeese, Modeling and guiding the creation of ethical human-AI teams, in: Proceedings of AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 469–479.

[28] B. Mittelstadt, Principles alone cannot guarantee ethical AI, Nat. Mach. Intell. 1 (11) (2019) 501–507.

[29] M. Serramia, M. Lopez-Sanchez, J.A. Rodriguez-Aguilar, M. Rodriguez, M. Wooldridge, J. Morales, C. Ansotegui, Moral values in norm decision making, in: Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems, 2018, pp. 1294–1302.

[30] J. Luo, J.C. Meyer, M. Knobbout, Reasoning about opportunistic propensity in multi-agent systems, in: Lecture Notes in Computer Science, vol. 10642, 2017, pp. 203–221.

[31] T.J.M. Bench-Capon, K. Atkinson, Abstract argumentation and values, in: Argumentation in Artificial Intelligence, 2009, pp. 45–64.

[32] P.L. Hill, D.K. Lapsley, Persons and situations in the moral domain, J. Res. Personal. 43 (2) (2009) 245–246, http://dx.doi.org/10.1016/j.jrp.2008.12.034.

[33] J. Brännmark, Moral disunitarianism, Philos. Q. 66 (264) (2015) 481–499, http://dx.doi.org/10.1093/pq/pqv114.

[34] I. Kola, R. Isufaj, C.M. Jonker, Does personalization help? Predicting how social situations affect personal values, in: HHAI2022: Augmenting Human Intellect, 2022, pp. 157–169, http://dx.doi.org/10.3233/FAIA220196.

[35] J. de Wet, D. Wetzelhütter, J. Bacher, Revisiting the trans-situationality of values in Schwartz's portrait values questionnaire, Qual. Quant. 53 (2) (2018) 685–711.

[36] C. Warren, A.P. Mcgraw, L. Van Boven, Values and preferences: Defining preference construction, Wiley Interdiscipl. Rev.: Cogn. Sci. 2 (2) (2011) 193–205.

[37] C. Schein, The importance of context in moral judgments, Perspect. Psychol. Sci. 15 (2) (2020) 207–215, http://dx.doi.org/10.1177/1745691620904083.

[38] R. Lera-Leri, F. Bistaffa, M. Serramia, M. Lopez-Sanchez, J. Rodriguez-Aguilar, Towards pluralistic value alignment: Aggregating value systems through ℓ_p-regression, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, 2022, pp. 780–788.

[39] E. Liscio, R. Lera-Leri, F. Bistaffa, R.I. Dobbe, C.M. Jonker, M. Lopez-Sanchez, J.A. Rodriguez-Aguilar, P.K. Murukannaiah, Value inference in sociotechnical systems, in: Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23, IFAAMAS, London, United Kingdom, 2023, pp. 1774–1780.

[40] N. Mouter, P. Koster, T. Dekker, Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments, Transp. Res. Part A: Policy Pract. 144 (2021) 54–73.

[41] A. Itten, N. Mouter, When digital mass participation meets citizen deliberation: Combining mini- and maxi-publics in climate policy-making, Sustainability 14 (8) (2022).

[42] European Values Study, 2021. https://europeanvaluesstudy.eu. Online.

[43] L.C. Siebert, E. Liscio, P.K. Murukannaiah, L. Kaptein, S.L. Spruit, J. van den Hoven, C.M. Jonker, Estimating value preferences in a hybrid participatory system, in: HHAI2022: Augmenting Human Intellect, IOS Press, Amsterdam, the Netherlands, 2022, pp. 114–127.

[44] J. González-Pachón, C. Romero, Bentham, Marx and Rawls ethical principles: In search for a compromise, Omega 62 (2016) 47–51.

[45] D. Adil, R. Peng, S. Sachdeva, Fast, provably convergent IRLS algorithm for p-norm linear regression, in: Proceedings of Advances in Neural Information Processing Systems, 2019, pp. 14189–14200.

[46] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[47] F. Brandt, V. Conitzer, U. Endriss, J. Lang, A. Procaccia, Handbook of Computational Social Choice, Cambridge University Press, 2016.

[48] Y. Chevaleyre, U. Endriss, J. Lang, N. Maudet, A short introduction to computational social choice, in: International Conference on Current Trends in Theory and Practice of Computer Science, Springer, 2007, pp. 51–69.

[49] E. Tsiporkova, V. Boeva, Multi-step ranking of alternatives in a multi-criteria and multi-expert decision making environment, Inform. Sci. 176 (18) (2006) 2673–2697.

[50] E. Herrera-Viedma, F. Herrera, F. Chiclana, A consensus model for multiperson decision making with different preference structures, IEEE Trans. Syst., Man, Cybern.-Part A: Syst. Hum. 32 (3) (2002) 394–402.

[51] C.L. Butler, A. Rothstein, On Conflict and Consensus: A Handbook on Formal Consensus Decision Making, Takoma Park, 2006.

[52] Z. Xu, X. Cai, Group consensus algorithms based on preference relations, Inform. Sci. 181 (1) (2011) 150–162.

[53] F.A. Ortega Riejos, M.Á. Pozo Montaño, J. Puerto Albandoz, Modelling and planning public cultural schedules for efficient use of resources, Comput. Oper. Res. 58 (2015) 9–23.

[54] J. Xiao, X. Wang, H. Zhang, Exploring the ordinal classifications of failure modes in the reliability management: An optimization-based consensus model with bounded confidences, Group Decis. Negot. 31 (1) (2022) 49–80.

[55] J. González-Pachón, C. Romero, Distance-based consensus methods: A goal programming approach, Omega 27 (3) (1999) 341–347.

[56] J. González-Pachón, C. Romero, Aggregation of ordinal and cardinal preferences: A framework based on distance functions, J. Multi-Criteria Decis. Anal. 15 (3–4) (2008) 79–85.

[57] J. González-Pachón, C. Romero, The design of socially optimal decisions in a consensus scenario, Omega 39 (2) (2011) 179–185.

[58] W.D. Cook, L.M. Seiford, Priority ranking and consensus formation, Manage. Sci. 24 (16) (1978) 1721–1732.

[59] W.D. Cook, L.M. Seiford, On the Borda-Kendall consensus method for priority ranking problems, Manage. Sci. 28 (6) (1982) 621–637.

[60] W.D. Cook, M. Kress, Ordinal Information and Preference Structures: Decision Models and Applications, Prentice-Hall, Inc., 1992.

[61] W.D. Cook, M. Kress, L.M. Seiford, A general framework for distance-based consensus in ordinal ranking models, European J. Oper. Res. 96 (2) (1997) 392–397.

[62] P.-L. Yu, A class of solutions for group decision problems, Manag. Sci. 19 (8) (1973) 936–946.

[63] P.-L. Yu, Multiple-Criteria Decision Making: Concepts, Techniques, and Extensions, vol. 30, Springer Science & Business Media, 2013.

[64] J. González-Pachón, C. Romero, Properties underlying a preference aggregator based on satisficing logic, Int. Trans. Oper. Res. 22 (2) (2015) 205–215.

[65] J. Bentham, An Introduction To the Principles of Morals and Legislation, Payne and Son, London, 1789.

[66] Y. Chevaleyre, P. Dunne, U. Endriss, J. Lang, M. Lemaitre, N. Maudet, J. Padget, S. Phelps, J. Rodrguez-Aguilar, P. Sousa, Issues in multiagent resource allocation, Informatica 30 (2006) 3–31.

[67] J. Rawls, A Theory of Justice, Oxford University Press, Oxford, 1973.

[68] I. van de Poel, L. Royakkers, Ethics, Technology, and Engineering: An Introduction, Wiley-Blackwell, Hoboken, NJ, 2011.

[69] D. Cooper, Value Pluralism and Ethical Choice, St. Martin Press, Inc., New York, 1993.

[70] C.A. Le Dantec, E.S. Poole, S.P. Wyche, Values as lived experience, in: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI '09, ACM Press, New York, 2009, p. 1141.

[71] A. Pommeranz, C. Detweiler, P. Wiggers, C.M. Jonker, Elicitation of situated values: Need for tools to help stakeholders and designers to reflect and communicate, Ethics Inf. Technol. 14 (4) (2012) 285–303, http://dx.doi.org/10.1007/s10676-011-9282-6.

[72] K.B. Francis, C. Howard, I.S. Howard, M. Gummerum, G. Ganis, G. Anderson, S. Terbeck, Virtual morality: Transitioning from moral judgment to moral action? in: X. Wan (Ed.), PLoS One 11 (10) (2016) 1–22, http://dx.doi.org/10.1371/journal.pone.0164374, URL https://dx.plos.org/10.1371/journal.pone.0164374.

[73] R.M. Chisholm, Supererogation and offence: A conceptual scheme for ethics, Ratio (Misc.) 5 (1) (1963).

[74] S. Scheffler, Valuing, in: Equality and Tradition: Questions of Value in Moral and Political Theory, first ed., Oxford University Press, 2012, p. 352.

[75] World Values Survey, 2021. https://www.worldvaluessurvey.org/wvs.jsp. Online.

[76] European Values Study, Integrated dataset, GESIS data archive, version 4.0.0, 2017, http://dx.doi.org/10.4232/1.13560, Online.

[77] F. Molteni, R. Ladini, F. Biolcati, A.M. Chiesi, G.M.D. Sani, S. Guglielmi, M. Maraffi, A. Pedrazzani, P. Segatti, C. Vezzoni, Searching for comfort in religion: Insecurity and religious behaviour during the COVID-19 pandemic in Italy, Eur. Soc. 23 (2021) S704–S720.

[78] C. Knill, C. Adam, S. Hurka, On the Road To Permissiveness? Change and Convergence of Moral Regulation in Europe, Oxford University Press, 2015.

[79] S. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, et al., Refining the theory of basic individual values, J. Personal. Soc. Psychol. 103 (4) (2012) 663.

[80] Wikipedia, Definition of permissive society, 2021, URL https://en.wikipedia.org/wiki/Permissive_society.

[81] O. Araque, L. Gatti, K. Kalimeri, MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction, Know.-Based Syst. 191 (C) (2020) http://dx.doi.org/10.1016/j.knosys.2019.105184.

[82] E. Liscio, A.E. Dondera, A. Geadau, C.M. Jonker, P.K. Murukannaiah, Cross-domain classification of moral values, in: Findings of the Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '22, ACL, Seattle, USA, 2022, pp. 2727–2745.

[83] G. Beliakov, H.B. Sola, T.C. Sánchez, A Practical Guide to Averaging Functions, vol. 329, Springer, 2016.