Towards Pluralistic Value Alignment: Aggregating Value Systems through ℓ_p-Regression

Roger Lera-Leri
IIIA-CSIC
Barcelona, Spain
rlera@iiia.csic.es

Filippo Bistaffa
IIIA-CSIC
Barcelona, Spain
filippo.bistaffa@iiia.csic.es

Marc Serramia
IIIA-CSIC
Barcelona, Spain
marcserr@iiia.csic.es

Maite Lopez-Sanchez
Universitat de Barcelona
Barcelona, Spain
maite_lopez@ub.edu

Juan Rodriguez-Aguilar
IIIA-CSIC
Barcelona, Spain
jar@iiia.csic.es

ABSTRACT

Dealing with the challenges of an interconnected globalised world requires to handle plurality. This is no exception when considering value-aligned intelligent systems, since the values to align with should capture this plurality. So far, most literature on value-alignment has just considered a single value system. Thus, this paper advances the state of the art by proposing a method for the aggregation of value systems. By exploiting recent results in the social choice literature, we formalise our aggregation problem as an optimisation problem. We then cast such problem as an *ℓ_p-regression problem*. By doing so, we provide a general theoretical framework to model and solve the above-mentioned 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 the European Values Study and we show how different *consensus value systems* can be obtained depending on the ethical principle of choice.

KEYWORDS

AI & Ethics; Value Systems; Optimisation

ACM Reference Format:
Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar. 2022. Towards Pluralistic Value Alignment: Aggregating Value Systems through ℓ_p-Regression. In *Proc.of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 9 pages.

1 INTRODUCTION

The new vision of human-centred Artificial Intelligence (AI) has spurred research on trustworthy, ethical AI that enhances human capabilities and empowers citizens and society to effectively deal with the challenges of an interconnected globalised world. Thus, developing trustworthy AI [8] that abides by human values has been a primary AI concern, as explicitly stated by the European Commission *Ethics Guidelines for Trustworthy AI* [13], the *Artificial Intelligence Act* [14], and the IEEE *Global Initiative on Ethics of Autonomous and Intelligent Systems* [22]. Within this vision, value alignment has raised as a core problem for AI. The goal of AI value alignment is to ensure that AI is properly aligned with human values [36]. Thus, designing an agent that aligns with values means that the agent does "what it morally ought to do, as defined by the individual or society" [16].

Thus, the problem of value alignment has spurred much research that tackles different dimensions of the problem such as value sensitive design [47], how to formalise the value alignment problem [41], how to identify values [24], how to learn value-aligned behaviours [1, 4, 30, 34, 35, 48], how to reason about values and act ethically [3], or how to align norms with values [29, 39]. A common assumption pervading 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 the society's. Nonetheless, as argued in [16] by following Rawls, humans hold a variety of reasonable but contrasting beliefs about values. Therefore, we live in a pluralistic world where people ascribe to different moral systems (value systems). Designing an AI system that aligns with a group of people with different moral systems poses a so-called *pluralistic value alignment* problem [16]. Indeed, this is the case, for instance, when conducting policy-making decisions that align with stakeholders with a variety of moral systems (e.g., [31, 32]), or when designing human-agent teams involving humans with also different moral systems [15]. Then, considering this value diversity, the following question arises: is there a way to decide what moral values an AI should align with? This is the research question that we address in this paper.

More precisely, in this paper we try to make headway in the pluralistic value alignment problem by addressing how to aggregate the different value systems of a group of individual entities (e.g., a set of stakeholders, a team, a country, or even a set of countries) to yield a *consensus* value system. To succeed in this endeavour, we identify two major challenges. First, as noted in [27], existing ethical codes are rather abstract and vague about moral values, hence not being specific enough to be action-guiding. This is also typically the case in the AI literature considering value systems (e.g., [5, 25, 29, 40]), which disregards a formal, concrete view of the concept of moral value. And yet, aggregating moral systems demands a formal notion of moral system. Second, from a social choice perspective, the aggregation of moral systems can be performed following different *ethical principles* (e.g., utilitarian or egalitarian). Therefore, it would be desirable to count on a general aggregation method that allows to set the ethical principle of choice.

Proc.of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

Given the above-mentioned challenges, here we follow an approach based on recent results in the social choice literature [20] to make the following novel contributions to the state of the art:

- We formally capture the notion of moral belief (or, as we name it, value interpretation) based on the work by Chisholm [11] in the Ethics literature. Our formal view of moral value interpretation enforces each individual to explicitly express its view on how values judge actions as well as its preferences over values. To express such preferences, we exploit the analytical framework in [20].
- Based on the social choice (distance) functions in [20], 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 (moral) 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 moral system can be cast as an ℓ_p*-regression problem* [2] (also called *norm approximation problem* [7]). 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 [2].
- We illustrate our approach with actual-world data from the European Values Study [43], showing how the disparity of interpretations of European citizens can be aggregated by means of our method. Furthermore, we fully characterise the impact of the ethical principle on the resulting consensus, hence providing useful insights for decision-makers concerned with obtaining a consensus of different value systems according to an ethical principle of choice.

The paper is organised as follows. Section 2 provides background on the distance functions that we require from the social choice literature. Section 3 formalises our notions of moral value interpretation and value system. 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 5.3. Finally, Section 6 reports on our empirical findings and Section 7 draws conclusions and sets paths to future research.

2 BACKGROUND: DISTANCE FUNCTIONS

This section introduces the background about the social choice functions that we employ in this paper to define our aggregation problems. We largely borrow from [20], where the authors define a generator of social choice functions (as a p-parameterised distance function) to obtain a consensus in a society. Such generator produces social functions that vary depending on a given principle (e.g., egalitarian, utilitarian, equity).[1] The general setting in [20] considers a society formed by n members ($i = 1, \ldots, n$). Each member of the society gives judgement values on m objects ($j, k = 1, \ldots, m$), which can be candidates, criteria, alternatives, etc. Furthermore:

- w_i is the weight (social influence) of the i-th member.
- $R_i[j,k]$ is the judgement value provided by the i-th member of the society when comparing the j-th and the k-th object. This judgement value can be cardinal, when the individual expresses some degree of preference over the alternatives, or ordinal, when they use a Boolean value to express their (crisp) preference for alternative j over k.
- $R_S[j,k]$ is the *consensus* value assigned by the society as a whole to the j-th object when compared with the k-th object. R_S is the unknown consensus that we seek to obtain.
- p is a metric parameter (i.e., a real number ≥ 1) that determines the *ethical principle* used to compute the consensus, in accordance with the terminology already established in the social choice literature [17–20].

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 [17–20]:

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

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

From the U_p distance function, the authors of [20] derive two particular cases of interest. First, by setting $p = 1$, the general distance in Equation 1 yields

$$U_B = \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1, k\neq j}^{m} w_i |R_i[j,k] - R_S[j,k]| \right]. \quad (2)$$

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 [6]) that maximises the total welfare.

By setting $p = \infty$, the distance function in Equation 1 yields

$$U_R = \max_{i,j,k} \left[|R_i[j,k] - R_S[j,k]| \right]. \quad (3)$$

Equation 3 represents the Chebyshev distance, which is equivalent to the weighted Minkowski distance for $p = \infty$. Finding the consensus in this case implies the minimisation of the disagreement of the member of the society most displaced with respect to the majority solution defined by the utilitarian case above (Equation 2). This solution is *egalitarian* [10] as 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 Rawl's principle [33]), leading to the point of maximum fairness. Note that, when considering the limit case $p \to \infty$, the weighting scheme in Equation 1 vanishes, hence the weight w_i does not appear in Equation 3.

In Section 4 we will employ the general distance function in Equation 1 to pose our problem of aggregating different value systems as that of computing a consensus. By leveraging such general distance function, Section 6 will consider both the social influence w_i and the ethical principle determined by p (including the utilitarian and egalitarian principles that stem from Equations 2 and 3) to study how the resulting consensus is affected.

3 MORAL VALUES AND VALUE SYSTEMS

Within Ethics, moral values (also called ethical principles) express the moral objectives *worth striving for* [45]. Examples of human

[1] 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 [20], it often results in an over-constrained optimisation problem that yields no solution.

values[2] include fairness, respect, freedom, security, or prosperity [9]. Every ethical theory considers one or more moral values that should guide our behaviour [12]. Therefore, ethical reasoning typically involves not a single moral value, but multiple moral values along with value preferences [5, 25, 39, 40], constituting a so-called *value system*. Value systems can be individual or shared by a society. In this section we focus on formalising the notion of value system to subsequently tackle the aggregation of multiple value systems in order to reach a *consensus* (or widely endorsed) value system.

We start by focusing on a formal view of the notion of moral value interpretation. Henceforth, we consider a set of agents Ag and a set of actions A. We also consider a set V of labels for moral values (e.g., "fairness", "respect"). Following Gabriel [16], here, we take the stance of values as guiding principles anchored in some set of evaluative judgements. However, as we live in a pluralistic world composed by a variety of reasonable and contrasting beliefs about value, Gabriel also defines values as placeholders for different *interpretations* of shared principles [16]. In this manner, we formally characterise moral values based on the relation between actions and values put forward by the Ethics literature. Indeed, as argued by Chisholm in [11], an action can be judged as either good or bad to perform (or to not perform) with respect to a given moral value. Thus, individuals from different cultural backgrounds might differently judge the same action with respect to the same moral value, and we refer to this 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 judge dress codes differently (e.g., favouring white).

Definition 3.1 (Moral value interpretation). Given a label $v \in V$ and a set of actions A, we say that an interpretation of v is a pair of action judgement functions (α_v^+, α_v^-) such that $\alpha_v^+, \alpha_v^- : A \to [-1, 1]$. Each of these functions takes an action and returns its evaluation with respect to the moral value label v. Function α_v^+ evaluates the praiseworthiness of performing an action, while $\alpha_v^-(a)$ evaluates the praiseworthiness of not performing an action. These evaluations are real numbers in the interval $[-1, 1]$: a positive number means that the moral value is being promoted, whereas a negative one stands for demotion.

Notice that judgement functions within a moral value interpretation allow an individual to quantify the moral praiseworthiness (right) and blameworthiness (wrong) of performing and not performing actions. Henceforth, given a moral value label $v \in V$, we shall say that the moral value interpretation of agent $i \in Ag$ of v is represented by the judgement functions $\alpha_{i,v}^+$ and $\alpha_{i,v}^-$. Finally, the following definition provides a compressed representation of the judgement functions of each agent $i \in Ag$.

Definition 3.2 (Judgement matrices). Given a set of labels of moral values V, an agent $i \in Ag$ and its interpretations of the moral values labels in V represented by $\alpha_{i,v}^+$ and $\alpha_{i,v}^-$, we define the judgement matrices $J_i^+, J_i^- \in \mathbb{R}^{|V| \times |A|}$ associated to agent i as

$$J_i = \left.\begin{matrix} J_i^+ = \begin{bmatrix} \alpha_{i,1}^+(a_1) & \cdots & \alpha_{i,1}^+(a_{|A|}) \\ \vdots & \ddots & \vdots \\ \alpha_{i,|V|}^+(a_1) & \cdots & \alpha_{i,|V|}^+(a_{|A|}) \end{bmatrix} \\ J_i^- = \begin{bmatrix} \alpha_{i,1}^-(a_1) & \cdots & \alpha_{i,1}^-(a_{|A|}) \\ \vdots & \ddots & \vdots \\ \alpha_{i,|V|}^-(a_1) & \cdots & \alpha_{i,|V|}^-(a_{|A|}) \end{bmatrix} \end{matrix}\right\}$$

The rows of these matrices are associated to the labels in V, whereas the columns are associated to the actions in A. For convenience, we denote the vertical concatenation of J_i^+ and J_i^- as J_i.

For simplicity, henceforth we will refer to a moral value interpretation of a given agent as the moral value of the agent.

Now we are ready to define the notion of value system for an individual agent. In short, a value system for some agent i will contain its moral value interpretations together with its preferences over moral values. Formally:

Definition 3.3 (Value system). Given a set of labels of moral values V and a set of available actions A, a value system $\mathcal{V}_i$ for $i \in Ag$ is

$$\mathcal{V}_i = \langle V, J_i, P_i \rangle$$

where J_i stands for the vertical concatenation of the judgement matrices J_i^+ and J_i^- of agent i and $P_i \in [0, 1]^{|V| \times |V|}$ is a matrix that contains the preferences over moral values for agent i. Thus, $P_i[j, k] \in [0, 1]$ is a graded preference provided by agent i when comparing the moral values corresponding to labels j and k (see Section 2), where 0.5 stands for indifference. For instance, an agent may prefer environmental protection over economic development with a grade of 0.75.

On the one hand, the literature has typically considered preferences over values in a qualitative manner: [5, 25] employ total orders, whereas [39, 40] try to be more flexible by resorting to partial orders and rankings respectively. Our definition of value system above follows an alternative, and more general, approach based on the work of [17–20], where preferences are expressed by *pairwise comparisons*. As discussed in Section 2, when using Boolean values in the preference matrix, we can express ordinal, qualitative preferences (like those in [5, 25, 39, 40]), whereas we can also express graded quantitative preferences through cardinal values. Furthermore, as argued in [20], the information in the preference matrix can be complete or incomplete.

On the other hand, notice that the value systems defined in [5, 39, 40] do not consider the link between values and actions. Moreover, although the approach in [25] does consider the relation between actions and values, it does not quantify it. Here, instead, our definition of value system is conceived to capture the alignment between actions and moral values.

4 FORMALISING THE AGGREGATION OF VALUE SYSTEMS

Having defined the concept of a value system $\mathcal{V}_i$ in Definition 3.3, in this section we tackle the problem of aggregating *multiple* value systems $\mathcal{V}_1, \ldots, \mathcal{V}_{|Ag|}$ (each associated to an agent $i \in Ag$)

[2] Sociology and Psychology have also extensively studied human values, which are often defined as abstract ideals that guide people's behaviour [37] or idealised standards with an "ought" character [26]. Social psychology also define values as ideals shared by members of a culture about what is good or bad [21].

and computing a value system $\mathcal{V}_S$ that *best* represents the aggregation of $\mathcal{V}_1, \ldots, \mathcal{V}_{|Ag|}$ according to a given *ethical principle* p.[3] Following the social choice literature discussed in Section 2, here we propose to cast this problem as the one of computing the value system $\mathcal{V}_S$ that minimises a certain distance measure defined along the lines of U_p (Equation 1). Being a value system composed by two separate components (i.e., the judgement functions J_i and the preferences over moral values P_i), in this paper we pose such a distance-minimisation problem as a two-step procedure, first to compute J_S (denoting the vertical concatenation of J_S^+ and J_S^-, according to Definition 3.2) and then to compute P_S.

Along these lines, by making use of Definition 3.2 we define the first distance function $U_p^{(J)}$ (representing the distance between the individual judgements J_i and the aggregated judgement J_S) as

$$U_p^{(J)} = \left[\sum_{i \in Ag} w_i \sum_{j=1}^{|V|} \sum_{k=1}^{|A|} \left| J_i^+[j,k] - J_S^+[j,k] \right|^p + \left| J_i^-[j,k] - J_S^-[j,k] \right|^p \right]^{1/p}. \quad (4)$$

In a similar way we define the second utility function $U_p^{(P)}$ (representing the distance between the individual preferences P_i and the aggregated preference P_S) as

$$U_p^{(P)} = \left[\sum_{i \in Ag} w_i \sum_{j=1}^{|V|} \sum_{k=1}^{|V|} \left| P_i[j,k] - P_S[j,k] \right|^p \right]^{1/p}. \quad (5)$$

Having defined our utility functions $U_p^{(J)}$ and $U_p^{(P)}$, we can now formally pose the *value system aggregation problem* as the problem of computing $\mathcal{V}_S = \langle V, J_S, P_S \rangle$ such that

$$J_S = \arg\min U_p^{(J)}, \quad (6)$$

$$P_S = \arg\min U_p^{(P)}. \quad (7)$$

In the following sections we will elaborate on how we solve the optimisation problems in Equations 6 and 7.

5 AN ℓ_p-REGRESSION APPROACH TO AGGREGATE VALUE SYSTEMS

In this section we show how Equation 6 and Equation 7 can be cast as ℓ_p-regression problems (also known as *norm approximation problems* [7]). Such a transformation incurs obvious computational benefits as it allows us to efficiently solve the above-mentioned problems for any p, as explained in Section 5.3.

5.1 Agreeing on moral values

We first prove Theorem 5.1, which shows how computing the aggregation of judgement functions (Equation 6) is equivalent to solving a properly defined ℓ_p-regression problem.

Theorem 5.1. *Computing the solution J_S of Equation 6 is equivalent to computing the solution x of the ℓ_p-regression problem*

$$\textit{minimise} \ \ \|Bx - b\|_p, \quad (8)$$

where $B \in \mathbb{R}^{2 \cdot |Ag| \cdot |V| \cdot |A| \times 2 \cdot |V| \cdot |A|}$ and $b \in \mathbb{R}^{2 \cdot |Ag| \cdot |V| \cdot |A|}$ are

$$B = \begin{bmatrix} w_1^{1/p} \cdot I \\ \vdots \\ w_{|Ag|}^{1/p} \cdot I \end{bmatrix}, \quad b = \begin{bmatrix} w_1^{1/p} \cdot \overrightarrow{J_1^+} \\ w_1^{1/p} \cdot \overrightarrow{J_1^-} \\ \vdots \\ w_{|Ag|}^{1/p} \cdot \overrightarrow{J_{|Ag|}^+} \\ w_{|Ag|}^{1/p} \cdot \overrightarrow{J_{|Ag|}^-} \end{bmatrix},$$

$I \in \mathbb{R}^{2 \cdot |V| \cdot |A| \times 2 \cdot |V| \cdot |A|}$ is the identity matrix of size $2 \cdot |V| \cdot |A|$, $\overrightarrow{(\cdot)}$ is the vectorisation *operation that turns a matrix into a 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 Equation 4 as

$$\left[\sum_{i \in Ag} w_i \sum_{h=1}^{|V| \cdot |A|} \left| \overrightarrow{J_i^+}[h] - \overrightarrow{J_S^+}[h] \right|^p + \left| \overrightarrow{J_i^-}[h] - \overrightarrow{J_S^-}[h] \right|^p \right]^{1/p} \quad (9)$$

and, subsequently, as

$$\left[\sum_{i \in Ag} \left\| w_i^{1/p} \cdot \overrightarrow{J_i^+} - w_i^{1/p} \cdot \overrightarrow{J_S^+} \right\|_p^p + \left\| w_i^{1/p} \cdot \overrightarrow{J_i^-} - w_i^{1/p} \cdot \overrightarrow{J_S^-} \right\|_p^p \right]^{1/p}. \quad (10)$$

To express Equation 10 as an ℓ_p-regression problem, we define $B \in \mathbb{R}^{2 \cdot |Ag| \cdot |V| \cdot |A| \times 2 \cdot |V| \cdot |A|}$ and $b \in \mathbb{R}^{2 \cdot |Ag| \cdot |V| \cdot |A|}$ as

$$B = \begin{bmatrix} w_1^{1/p} \cdot I \\ \vdots \\ w_{|Ag|}^{1/p} \cdot I \end{bmatrix}, \quad b = \begin{bmatrix} w_1^{1/p} \cdot \overrightarrow{J_1^+} \\ w_1^{1/p} \cdot \overrightarrow{J_1^-} \\ \vdots \\ w_{|Ag|}^{1/p} \cdot \overrightarrow{J_{|Ag|}^+} \\ w_{|Ag|}^{1/p} \cdot \overrightarrow{J_{|Ag|}^-} \end{bmatrix}.$$

We can finally formulate Equation 6 as

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

The solution of the above-defined problem (i.e., the vector x) is J_S, i.e., the vertical concatenation of $\overrightarrow{J_S^+}$ and $\overrightarrow{J_S^-}$. □

5.2 Aggregating preferences over moral values

Having tackled Equation 6, in the same vein we prove Theorem 5.2, which shows how computing the aggregation of preferences (Equation 7) is also equivalent to an ℓ_p-regression problem.

Theorem 5.2. *Computing the solution P_S of Equation 7 is equivalent to computing the solution x of the ℓ_p-regression problem*

$$\textit{minimise} \ \ \|Cx - c\|_p, \quad (11)$$

where $C \in \mathbb{R}^{|Ag| \cdot |V|^2 \times |V|^2}$ and $c \in \mathbb{R}^{|Ag| \cdot |V|^2}$ are defined as

$$C = \begin{bmatrix} w_1^{1/p} \cdot I \\ \vdots \\ w_{|Ag|}^{1/p} \cdot I \end{bmatrix}, \quad c = \begin{bmatrix} w_1^{1/p} \cdot \overrightarrow{P_1} \\ \vdots \\ w_{|Ag|}^{1/p} \cdot \overrightarrow{P_{|Ag|}} \end{bmatrix},$$

and $I \in \mathbb{R}^{|V|^2 \times |V|^2}$ is the identity matrix of size $|V|^2$.

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

Proof. By following a procedure similar to the one adopted for Theorem 5.1, we rewrite Equation 5 as

$$\left[\sum_{i\in Ag}\left\|w_i^{1/p}\cdot\overrightarrow{P_i}-w_i^{1/p}\cdot\overrightarrow{P_S}\right\|_p^p\right]^{1/p}. \tag{12}$$

To express Equation 12 as an ℓ_p-regression problem, we define $C\in\mathbb{R}^{|Ag|\cdot|V|^2\times|V|^2}$ and $c\in\mathbb{R}^{|Ag|\cdot|V|^2}$ as

$$C=\begin{bmatrix} w_1^{1/p}\cdot I\\ \vdots\\ w_{|Ag|}^{1/p}\cdot I\end{bmatrix},\quad c=\begin{bmatrix} w_1^{1/p}\cdot\overrightarrow{P_1}\\ \vdots\\ w_{|Ag|}^{1/p}\cdot\overrightarrow{P_{|Ag|}}\end{bmatrix}.$$

We can finally formulate Equation 6 as

$$\text{minimise}\ \ \|Cx-c\|_p.$$

The solution of such a problem (i.e., the vector x) is $\overrightarrow{P_S}$. □

5.3 Solving the ℓ_p-regression problem

We now discuss the computational aspects of solving Equation 8. We remark that the following discussion also applies to Equation 11, since both are ℓ_p-regression problems.

For $p=1$, Equation 8 represents an *absolute residuals approximation problem*. For $p=\infty$, we are dealing with a *Chebyshev approximation problem* or, in other words, a *Min-Max approximation problem*. In both cases, the solution to Equation 8 can be computed by means of *Linear Programming* [7]. For $p=2$, Equation 8 can be solved analytically by treating it as *Least Squares* problem,[4] whose optimal solution is

$$x=(B^TB)^{-1}B^Tb. \tag{13}$$

We employ this well-known result in next Proposition 5.3, where we show that, for $p=2$, the judgement matrices of the aggregated value system can be elegantly obtained as the *weighted arithmetic mean* of the individual judgement matrices. An equivalent result can be easily obtained for the preference matrix of the aggregated value system (proof not included here for the sake of conciseness).

Proposition 5.3. *For $p=2$, J_S can be analytically computed as the* weighted arithmetic mean *of $J_1,\dots,J_{|Ag|}$, where the weights are $w_1,\dots,w_{|Ag|}$.*

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

$$\begin{aligned}(B^TB)^{-1}&=\left(\begin{bmatrix} w_1^{1/2}\cdot I & \cdots & w_{|Ag|}^{1/2}\cdot I\end{bmatrix}\begin{bmatrix} w_1^{1/2}\cdot I\\ \vdots\\ w_{|Ag|}^{1/2}\cdot I\end{bmatrix}\right)^{-1}\\ &=\left(\sum_{i=1}^{|Ag|}w_i\cdot I\right)^{-1}=\left(\sum_{i=1}^{|Ag|}w_i\right)^{-1}I.\end{aligned} \tag{14}$$

Notice that Equation 14 is a diagonal matrix in which the elements of the diagonal are all equal to the inverse of the sum of the weights.

[4] 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 [7].

By making use of the above result, we explicitly compute x as

$$\begin{aligned}x&=\left(\sum_{i=1}^{|Ag|}w_i\right)^{-1}\overbrace{\begin{bmatrix} w_1^{1/2}\cdot I & \cdots & w_{|Ag|}^{1/2}\cdot I\end{bmatrix}}^{B^T}\begin{bmatrix} w_1^{1/2}\cdot\overrightarrow{J_1^+}\\ w_1^{1/2}\cdot\overrightarrow{J_1^-}\\ \vdots\\ w_{|Ag|}^{1/2}\cdot\overrightarrow{J_{|Ag|}^+}\\ w_{|Ag|}^{1/2}\cdot\overrightarrow{J_{|Ag|}^-}\end{bmatrix}\\ &=\frac{\sum_{i=1}^{|Ag|}w_i\cdot\begin{bmatrix}\overrightarrow{J_i^+}\\ \overrightarrow{J_i^-}\end{bmatrix}}{\sum_{i=1}^{|Ag|}w_i}=\frac{\sum_{i=1}^{|Ag|}w_i\cdot\overrightarrow{J_i}}{\sum_{i=1}^{|Ag|}w_i}.\end{aligned} \tag{15}$$

Thus, each element of x (i.e., of J_S) is the weighted mean of the corresponding elements of J_i, according to the weights $w_1,\dots,w_{|Ag|}$. □

For any $p\notin\{1,2,\infty\}$, Equation 6 technically represents a non-linear problem. Nonetheless, by exploiting the structure of Equation 8 as an ℓ_p-regression problem, we are able to overcome such a computational challenge and solve it for any p. Our solution algorithm of choice is the state of the art algorithm called *Iteratively Reweighted Least Squares* (IRLS) [2], the only available approach for ℓ_p-regression that is guaranteed to converge for any value of p.[5]

6 CASE STUDY: EUROPEAN VALUES STUDY

The European Values Study (EVS) [43] is a large-scale survey research programme on European values. It collaborates with the World Values Survey [44] and it aims at providing data about the variety of positions that citizens from different European countries have regarding basic values such as well-being, solidarity, or democracy. Although the EVS survey covers a wide range of questions and values, here we just focus on a couple of values and three questions for 34 different European countries. Indeed, we remark that, far from aiming at providing realistic insights about European values, our study only aims at illustrating how our aggregation approach works with a simplified example.

Specifically, Section 6.1 details how we obtain the value systems of the considered European countries from the EVS data. Next, Section 6.2 aims at discussing the impacts of the parameters of our aggregation approach (namely, the ethical principle p and the social influence w_i) on the resulting value system. Finally, Section 6.3 characterises the space of ethical principles, providing useful insights for decision-makers concerned with obtaining a consensus of different value systems.

6.1 Value systems characterisation

We resort to the EVS data in [42] to create the value system of each country. Specifically, here we consider two moral values: *religiosity* (*rl*) and *permissiveness* (*pr*).[6] We then proceed by characterising these two moral values in terms of their value judgement functions.

[5] Our code is available at https://github.com/RogerXLera/ValueSystemsAggregation. The code by the authors of [2] is available at https://github.com/fast-algos/pIRLS.

[6] Although these values of *religiosity* [28] and *permissiveness* [23] can be related to those of *tradition* and *tolerance* from the Schwartz's revised model of values [38], we choose them to better fit EVS's data. In fact, one may even think that secularism seems a better alternative to permissiveness when comparing it to religiosity. However, we argue that permissiveness [23] is better suited, as it is specifically related to sexual freedom [46], and the data from EVS we use relates to homosexual couples and divorce.

Table 1: Value preferences (columns 2–3) and value judgement functions (columns 4–7) for the countries in the EVS.

Country	$P_i[rl, pr]$	$P_i[pr, rl]$	$\alpha^+_{rl}(ad)$	$\alpha^+_{rl}(dv)$	$\alpha^+_{pr}(ad)$	$\alpha^+_{pr}(dv)$
AL	0.72	0.28	−0.64	−0.33	−0.58	−0.18
AM	0.83	0.17	−0.73	−0.48	−0.71	−0.45
AT	0.44	0.56	0.06	0.19	0.32	0.54
AZ	0.60	0.40	−0.83	−0.32	−0.85	−0.13
BA	0.81	0.19	−0.46	−0.09	−0.31	0.15
BG	0.61	0.39	−0.42	−0.05	−0.32	0.16
BY	0.56	0.44	−0.67	−0.04	−0.52	0.20
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
HU	0.46	0.54	−0.27	0.00	−0.10	0.25
IS	0.40	0.60	0.54	0.41	0.70	0.71
IT	0.66	0.34	−0.15	0.12	0.05	0.51
LT	0.48	0.52	−0.56	0.03	−0.50	0.21
ME	0.87	0.13	−0.45	−0.34	−0.26	−0.15
MK	0.81	0.19	−0.52	−0.09	−0.30	0.27
NL	0.32	0.68	0.26	0.19	0.52	0.63
NO	0.35	0.65	0.33	0.38	0.64	0.68
PL	0.80	0.20	−0.49	−0.17	−0.20	0.34
PT	0.67	0.33	0.11	0.02	0.19	0.35
RO	0.82	0.18	−0.62	−0.25	−0.29	0.00
RS	0.75	0.25	−0.62	−0.09	−0.51	0.21
RU	0.51	0.49	−0.59	0.08	−0.56	0.20
SE	0.31	0.69	0.44	0.56	0.61	0.76
SI	0.36	0.64	−0.22	0.16	0.10	0.53
SK	0.58	0.42	−0.53	−0.05	−0.31	0.31

For simplicity, we consider the judgement of two actions: *adoption by homosexual couples* (*ad*) and *divorcing* (*dv*).

In order to characterise Europeans' position on the religiosity value, 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 that consider religion important from the ones who do not. Columns 2–3 of Table 1 show the respective percentages per country, which we also interpret as the degree of preference of each value. Formally, we denote the preference degree of value rl over pr in country i as $P_i[rl, pr]$, in line with the notation of $P_i[j, k]$ in Equation 5. 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 adoption and divorce (columns 4–5 of Table 1). Similarly, we obtain the judgements of non-religious people (columns 6–7 of Table 1). Since EVS lacks specific questions regarding the non-performance of actions, we assume that $\alpha^-_v(a) = -\alpha^+_v(a)$. Intuitively, if performing a is detrimental for v (i.e., $\alpha^+_v(a) < 0$), we interpret not performing a as promoting v (i.e., $\alpha^-_v(a) > 0$).

6.2 Aggregating European value systems

Once the value system for each European country has been specified, we are now ready to aggregate them into a consensus value system. This section is devoted to discuss how the resulting value system is affected by the parameters of our aggregation approach. Specifically, we focus on the impacts of the ethical principle p and the social influence w_i of each country.

To this end, we consider $p = 1$ (fully utilitarian) and $p = \infty$ (fully egalitarian). In addition, we consider another critical value of p (denoted as $\bar{p}$) that marks the "transition point" from the set of ethical principles that lean towards the fully utilitarian one and the ones that lean towards the fully egalitarian one. We provide a more detailed discussion on $\bar{p}$ and on the characterisation of the space of ethical principles in Section 6.3.

As for social influence, we consider two cases: a first one where all countries have the same social influence ($w_i = 1$) and a second where social influence w_i is set according to the population of each country. Specifically, for this second case we set w_i to be the population of country i normalised according to the total population of all considered countries (i.e., $w_i = \text{population}_i / \sum_j \text{population}_j$).[7]

[7] The population data is obtained from the *Worldometers* website (https://www.worldometers.info/world-population/population-by-country, accessed in Sept. 2021). The total population of the considered countries is approximately 687 million people.

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

p	$P_S[rl, pr]$	$P_S[pr, rl]$	$\alpha^+_{rl}(ad)$	$\alpha^+_{rl}(dv)$	$\alpha^+_{pr}(ad)$	$\alpha^+_{pr}(dv)$
1	0.504	0.496	−0.288	0.013	−0.054	0.318
$\bar{p}$	0.543	0.457	−0.208	0.030	−0.029	0.295
∞	0.580	0.420	−0.158	0.029	−0.077	0.234

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

p	$P_S[rl, pr]$	$P_S[pr, rl]$	$\alpha^+_{rl}(ad)$	$\alpha^+_{rl}(dv)$	$\alpha^+_{pr}(ad)$	$\alpha^+_{pr}(dv)$
1	0.444	0.556	0.007	0.123	0.187	0.503
$\bar{p}$	0.512	0.488	−0.132	0.108	0.053	0.396
∞	0.580	0.420	−0.158	0.029	−0.077	0.234

In our case study $|Ag| = 34$, $|V| = 2$, and $|A| = 2$, hence the problem in Equation 8 involves a matrix B of size 272×8 and a vector b of size 272. On the other hand, the problem in Equation 11 involves a matrix C of size 136×4 and a vector c of size 136.

6.2.1 Discussion of the results. Table 2 and 3 show the results of the aggregation with equal social influence and with social influence considering the population, respectively. In the first case (Table 2), w_i is set to 1 for all countries and the *consensus European value system* asserts that religiosity is generally preferred over permissiveness, since we obtain an aggregated preference $P_S[rl, pr] > 0.5$ for all values of p. However, we notice that for $p = 1$ such 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 p. This transition towards a greater value of $P_S[rl, pr]$ is due to the fact that, for greater values of p, the consensus tends to reduce the maximum disagreement with respect to the most "extreme" preference, which in our case is the one of Georgia ($P_i[rl, pr] = 0.94$). In addition, we notice that for all values of p the adoption by homosexual couples is valued negatively, whereas divorce is slightly accepted.

Conversely, when considering w_i to be determined by the population (Table 3), we observe significant changes with respect to the previous case. Specifically, for $p = 1$ (i.e., fully utilitarian ethical principle) we observe that permissiveness is preferred over religiosity (since $P_S[pr, rl]$ is larger than $P_S[rl, pr]$) and that both adoption by homosexual couples and divorce can be accepted. This shifting in the consensus is due to the social influence that wield highly populated countries in Europe —such as Germany, France, or Great Britain— which happen to prefer permissiveness over religiosity. Nonetheless, as we increase p we notice the same transition towards religiosity observed in Table 2. Indeed, we notice that for $p = \infty$ we obtain the same results for the weighted (varying w_i) and unweighted (equal w_i) cases, confirming that weights w_i vanish when considering the $\|\cdot\|_\infty$ norm in accordance with Equation 3.

Overall, choosing the aggregation parameters (i.e., p and w_i) amounts to deciding between swaying towards prevailing value systems that represent the majority or towards value systems that lay closer to divergent opinions.

6.3 Characterising the space of ethical principles

In this section we aim at characterising the space of ethical principles used in the aggregation of the European value systems previously described. We do so in order to determine whether a given p produces a consensus leaning towards the utilitarian ($p = 1$) or the egalitarian ($p = \infty$) ethical principle. To achieve this objective, we compute the aggregated judgement matrix J_S considering a given p (denoted as $J_S^{(p)}$) and we measure the distance between $J_S^{(p)}$ and the one corresponding to $p = 1$ and $p = \infty$, denoted as $J_S^{(1)}$ and $J_S^{(\infty)}$ respectively. Formally, we denote these two distances as $\left\|J_S^{(1)} - J_S^{(p)}\right\|_p$ and $\left\|J_S^{(p)} - J_S^{(\infty)}\right\|_p$. In an equivalent way, we define $\left\|P_S^{(1)} - P_S^{(p)}\right\|_p$ and $\left\|P_S^{(p)} - P_S^{(\infty)}\right\|_p$ for value preferences.

By making use of the above-defined distances, we can determine a *transition point* (denoted as $\bar{p}$) that is equidistant from the fully utilitarian and fully egalitarian consensuses. Along these lines, we can characterise an *utilitarian zone* composed by all $p < \bar{p}$, i.e., all ethical principles leaning towards the fully utilitarian case. Similarly, we can characterise an *egalitarian zone* composed by all $p > \bar{p}$, i.e., all ethical principles leaning towards the fully egalitarian case. Finally, we also characterise a *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 ($p = \infty$). Formally, we define the *fully egalitarian zone* as the set of all $p > \hat{p}$ such that

$$\frac{\left\|J_S^{(p)} - J_S^{(\infty)}\right\|_p}{\left\|J_S^{(1)} - J_S^{(\infty)}\right\|_1} < \epsilon.$$

Figure 1 plots the distances between the consensus judgement matrix ($J_S^{(p)}$) and the consensus value preference matrix ($P_S^{(p)}$) with respect to the fully utilitarian ($J_S^{(1)}$ and $P_S^{(1)}$) and fully egalitarian consensus matrices ($J_S^{(\infty)}$ and $P_S^{(\infty)}$) as the value of the ethical principle p increases. The two graphs on the top row of the figure refer to the case where all countries have equal social influence, whereas the two graphs on the bottom row refer to the case where social influence is determined by the population.

We notice that both the transition and the limit point are different when aggregating judgements (left) and value preferences (right). In other words, a given p can be interpreted as "more utilitarian" or "more egalitarian" depending on the actual position of the transition point. As an example, let us consider $p = 2.5$ in the weighted aggregation case (bottom row). Such a p can be intended as egalitarian when aggregating judgements, but slightly utilitarian when aggregating value preferences. Henceforth, the ability of being able to select *separate* values of p for judgements and preferences is fundamental to compute a consensus value system reflecting a particular ethical principle (e.g., egalitarian), hence motivating our choice of tackling the value system aggregation as a two-step procedure (Equation 6 and Equation 7).

In conclusion, the visual analysis displayed in Figure 1 is meant to provide useful guidance for decision-makers concerned with obtaining a consensus of different value systems following an ethical principle of choice. In general, a decision maker must start by plotting the distance between consensuses as we do in Figure 1. Thereafter, they can characterise the *common*, for both judgement and value preference, utilitarian and egalitarian zones. Thus, $[1, \bar{p}_m)$, where $\bar{p}_m = min(\bar{p}_J, \bar{p}_P)$, defines the utilitarian zone, whereas $(\bar{p}_M, \infty)$, where $\bar{p}_M = max(\bar{p}_J, \bar{p}_P)$, defines the egalitarian zone. Notice that $[\bar{p}_m, \bar{p}_M]$ contains ethical principles that lie in different zones for judgement and preferences (e.g., this is the case for the above-mentioned example considering $p = 2.5$ in the weighted aggregation case). Once characterised the common utilitarian and egalitarian zones, the decision maker is ready to choose a value for the ethical principle p.

7 CONCLUSIONS

In this paper we considered the pluralistic value alignment problem, contributing to the state of the art in the following ways.

We tackled this problem by proposing a framework for the formalisation and aggregation of different moral value systems. By rooting our framework in the social choice literature, we cast the

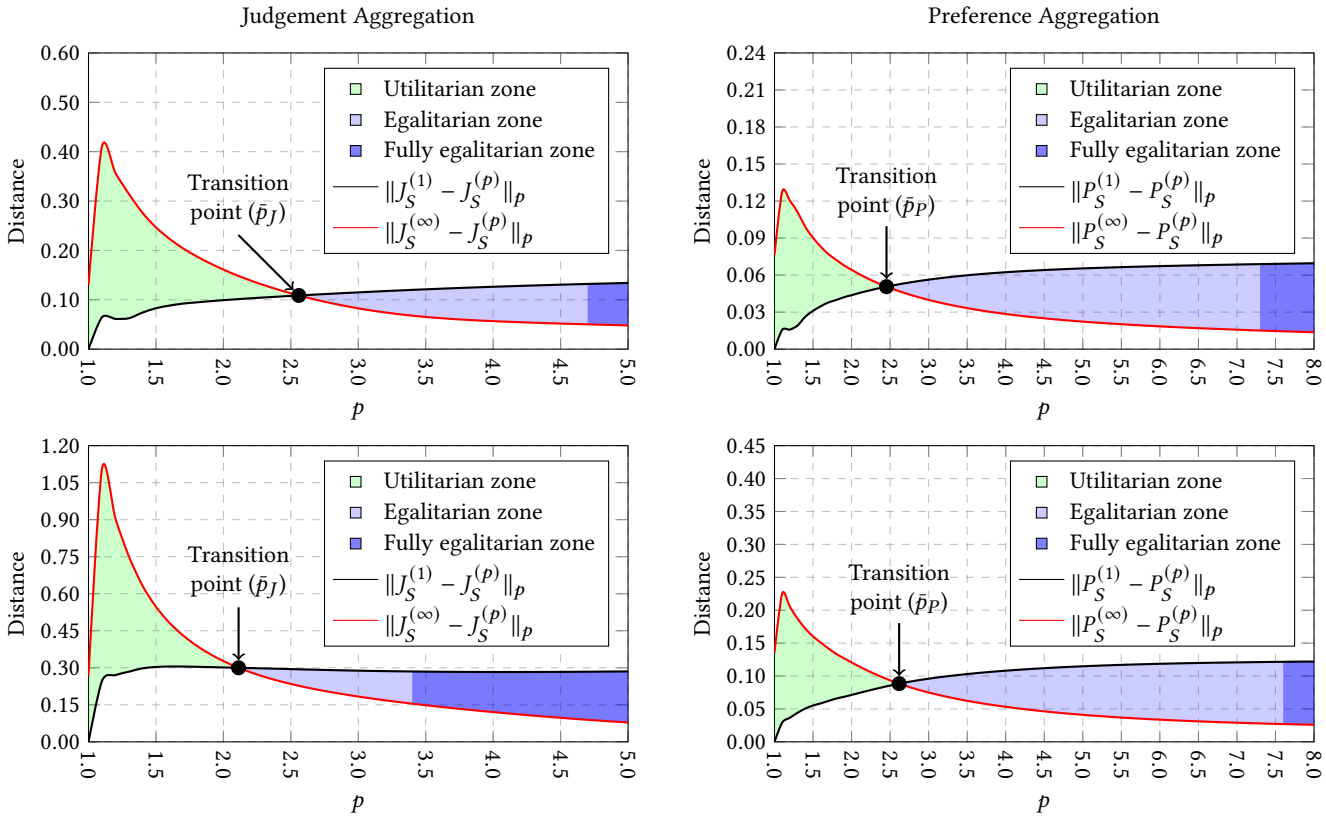

Figure 1: Distance between the consensus (J_S on the left and P_S on the right) computed according to ethical principle p and the the consensuses computed according to $p = 1$ (fully utilitarian, black line) and $p = \infty$ (fully egalitarian, red line). The transition point $\bar{p}$ is the ethical principle producing a consensus equidistant from the fully utilitarian and fully egalitarian ones. Hence, $\bar{p}$ 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 that is approximately equal (to a small ϵ) to the fully egalitarian one. Equal social influence on top, social influence according to the population at the bottom.

value system aggregation problem —where preferences and opinions over the moral values within the value system are uttered by individuals— as an ℓ_p-regression problem. Nonetheless, we also considered the Ethics literature to formalise the concepts of moral value interpretation through action judgement functions. Along these lines, our proposal follows a two-step process: first, we compute the aggregation of moral value interpretations; and subsequently, we aggregate the preferences over moral values. Moreover, the proposed framework is general enough to allow for the exploration of different ethical principles —which vary from utilitarian (maximum utility) to egalitarian (maximum fairness)— and that can also consider the social influence of the individuals.

We illustrated how different consensus value systems can result when applying our framework to the aggregation of 34 alternative European value systems. Overall, we highlighted that, when aggregating different value systems, the choice of aggregation parameters amounts to deciding between swaying towards prevailing value systems that represent the majority or towards value systems that lay closer to divergent opinions.

As future work we envision two research paths. First, we plan to generalise our framework so that each member i of the society Ag can choose their own ethical principle p_i when aggregating value systems. This might require the development of new social choice functions as well as new computational tools to account for multiple ethical principles. Second, notice that in this paper we have considered a particular family of social choice functions. However, we believe that the pluralistic value alignment problem opens an interesting research path for research on social choice, since value systems are complex objects to aggregate. Therefore, we plan to investigate further social choice functions to aggregate value systems and study their properties.

ACKNOWLEDGMENTS

This work was funded by projects Crowd4SDG (H2020-872944), TAILOR (H2020-952215), COREDEM (H2020-785907), and 21S01802-001 from Barcelona City Council through the Fundació Solidaritat de la UB. Grant PID2019-104156GB-I00 funded by MCIN/AEI 10.13039/501100011033.

REFERENCES

[1] David Abel, James MacGlashan, and Michael L Littman. 2016. Reinforcement Learning as a Framework for Ethical Decision Making. In *AAAI Workshop: AI, Ethics, and Society*.

[2] Deeksha Adil, Richard Peng, and Sushant Sachdeva. 2019. Fast, Provably Convergent IRLS Algorithm for p-Norm Linear Regression. In *Proceedings of Advances in Neural Information Processing Systems*. 14189–14200.

[3] Nirav Ajmeri. 2018. *Engineering Multi-Agent Systems for Ethics and Privacy-Aware Social Computing*. Ph.D. Dissertation. North Carolina State University.

[4] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. 2019. Incorporating Behavioral Constraints in Online AI Systems. *Proceedings of AAAI Conference on Artificial Intelligence*, 3–11.

[5] Trevor J. M. Bench-Capon and Katie Atkinson. 2009. Abstract Argumentation and Values. In *Argumentation in Artificial Intelligence*. 45–64.

[6] Jeremy Bentham. 1789. *An Introduction to the Principles of Morals and Legislation*. Payne and Son, London.

[7] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

[8] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. 2021. Trustworthy AI. In *Reflections on Artificial Intelligence for Humanity*. 13–39.

[9] An-Shou Cheng and Kenneth R Fleischmann. 2010. Developing a Meta-Inventory of Human Values. *Proceedings of American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[10] Yann Chevaleyre, Paul Dunne, Ulle Endriss, Jerome Lang, Michel Lemaitre, Nicolas Maudet, Julian Padget, Steve Phelps, Juan A Rodrguez-Aguilar, and Paulo Sousa. 2006. Issues in Multiagent Resource Allocation. *Informatica* 30 (2006), 3–31.

[11] Roderick M. Chisholm. 1963. Supererogation and Offence: A Conceptual Scheme for Ethics. *Ratio (Misc.)* 5, 1 (1963).

[12] David Cooper. 1993. *Value Pluralism and Ethical Choice*. St. Martin Press, Inc., New York.

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

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

[15] Christopher Flathmann, Beau G Schelble, Rui Zhang, and Nathan J McNeese. 2021. Modeling and Guiding the Creation of Ethical Human-AI Teams. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*. 469–479.

[16] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (2020), 411–437.

[17] Jacinto González-Pachón and Carlos Romero. 1999. Distance-Based Consensus Methods: a Goal Programming Approach. *Omega* 27, 3 (1999), 341–347.

[18] Jacinto González-Pachón and Carlos Romero. 2008. Aggregation of Ordinal and Cardinal Preferences: a Framework Based on Distance Functions. *Journal of Multi-Criteria Decision Analysis* 15, 3-4 (2008), 79–85.

[19] Jacinto González-Pachón and Carlos Romero. 2011. The Design of Socially Optimal Decisions in a Consensus Scenario. *Omega* 39, 2 (2011), 179–185.

[20] Jacinto González-Pachón and Carlos Romero. 2016. Bentham, Marx and Rawls Ethical Principles: In Search for a Compromise. *Omega* 62 (2016), 47–51.

[21] Jonathan Haidt. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Vintage.

[22] Institute of Electrical and Electronics Engineers. 2019. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html. Online.

[23] Christoph Knill, Christian Adam, and Steffen Hurka. 2015. *On the Road to Permissiveness? Change and Convergence of Moral Regulation in Europe*. Oxford University Press, USA.

[24] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proceedings of International Conference on Autonomous Agents and MultiAgent Systems*. 799–808.

[25] Jieting Luo, John-Jules Ch. Meyer, and Max Knobbout. 2017. Reasoning About Opportunistic Propensity in Multi-Agent Systems. In *Lecture Notes in Computer Science*, Vol. 10642. 203–221.

[26] Gregory Maio. 2016. *The Psychology of Human Values*. Routledge.

[27] Brent Mittelstadt. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.

[28] Francesco Molteni, Riccardo Ladini, Ferruccio Biolcati, Antonio M. Chiesi, Giulia Maria Dotti Sani, Simona Guglielmi, Marco Maraffi, Andrea Pedrazzani, Paolo Segatti, and Cristiano Vezzoni. 2021. Searching for Comfort in Religion: Insecurity and Religious Behaviour During the COVID-19 Pandemic in Italy. *European Societies* 23 (2021), S704–S720.

[29] Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems*. 907–915.

[30] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Ramazon Kush, Murray Campbell, Moninder Singh, and Francesca Rossi. 2019. Teaching AI Agents Ethical Values Using Reinforcement Learning and Policy Orchestration. *IBM Journal of Research and Development* 63, 4/5 (2019), 6377–6381.

[31] Klara Pigmans, Huib Aldewereld, Virginia Dignum, and Neelke Doorn. 2019. The Role of Value Deliberation to Improve Stakeholder Participation in Issues of Water Governance. *Water Resources Management* 33, 12 (2019), 4067–4085.

[32] Klara Pigmans, Neelke Doorn, Huib Aldewereld, and Virginia Dignum. 2017. Decision-Making in Water Governance: From Conflicting Interests to Shared Values. In *Responsible Innovation*. Springer, 165–178.

[33] John Rawls. 1973. *A Theory of Justice*. Oxford University Press, Oxford.

[34] Mark O. Riedl and B. Harrison. 2016. Using Stories to Teach Human Values to Artificial Agents. In *Proceedings of AAAI Workshop: AI, Ethics, and Society*.

[35] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez Aguilar. 2021. Multi-Objective Reinforcement Learning for Designing Ethical Environments. In *Proceedings of International Joint Conference on Artificial Intelligence*. 545–551.

[36] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

[37] Shalom Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture* 2, 1 (2012), 2307–0919.

[38] Shalom Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the Theory of Basic Individual Values. *Journal of Personality and Social Psychology* 103, 4 (2012), 663.

[39] Marc Serramia, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2020. A Qualitative Approach to Composing Value-Aligned Norm Systems. In *Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems*. 1233–1241.

[40] Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Manel Rodriguez, Michael Wooldridge, Javier Morales, and Carlos Ansotegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems*. 1294–1302.

[41] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perello-Moragues. 2019. Value Alignment: a Formal Approach. In *Proceedings of Responsible Artificial Intelligence Agents Workshop*.

[42] European Values Study. 2017. Integrated Dataset, GESIS Data Archive, Version 4.0.0. Online. https://doi.org/10.4232/1.13560

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

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

[45] Ibo van de Poel and Lambèr Royakkers. 2011. *Ethics, Technology, and Engineering: An Introduction*. Wiley-Blackwell, Hoboken, NJ.

[46] Wikipedia. 2021. Definition of Permissive Society. https://en.wikipedia.org/wiki/Permissive_society.

[47] Till Winkler and Sarah Spiekermann. 2021. Twenty Years of Value Sensitive Design: a Review of Methodological Practices in VSD Projects. *Ethics and Information Technology* 23, 1 (2021), 17–21.

[48] Yueh-Hua Wu and Shou-De Lin. 2018. A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents. In *Proceedings of AAAI Conference on Artificial Intelligence*. 1687–1694.