# Hierarchical Consensus Value Aggregation for Value-Aligned Multi-Agent Systems

Joseph Trevorrow[0000-0002-8668-1454] and Nirav Ajmeri[0000-0003-3627-097X]

University of Bristol, Bristol, BS8 1UB, UK
j.trevorrow@bristol.ac.uk

**Abstract.** Value aggregation is a method of group decision-making that combines agents' personal values to ensure that decisions align to the beliefs and preferences of the group. Personal values represent an agent's individual beliefs regarding what is important in their lives, motivating action. The output of value aggregation is a set of personal values representative of all agents, found according to a pre-defined ethical principle. An ethical principle reflects beliefs about how the group should be organised, and the standards that guide collective decision-making.
Current methods overlook agents' preferences for different ethical principles, resulting in a consensus that may not accurately reflect the diverse perspectives within the group. To address this, we introduce Hierarchical Consensus Value Aggregation (HCVA), a novel method that elicits agents' preferences of ethical principles. HCVA finds a representative ethical principle, which it uses for aggregation of agents' personal values. Using real-world data from the European Social Survey (ESS), we compare HCVA to competing aggregation techniques that are (1) fixed to a singular ethical principle and (2) able to represent different ethical principles at the same time. Results show HCVA is better in representative and fairness metrics in situations where an agent's principle preferences are representative of personal values. When an agent's principle preferences are included in metrics, HCVA performs better than selected baselines.

**Keywords:** Value-Alignment, Multi-Agent Systems, Group Decision-Making

## 1 Introduction

Group decision-making tools are important when agents hold different preferences about what makes a good decision. In group decision making, *value aggregation* serves as a method for determining appropriate actions to take considering agents' *personal values*. Personal values are defined as what matters to an individual in life or within a set context [7, 29, 25], indicating an agent's desirable goals and preferred outcomes, motivating action. The output of value aggregation is a consensus set of personal values that can be representatively used for ethical reasoning as a group [9, 11, 24]. Current value aggregation methods [11, 16, 24] rely on a pre-defined ethical principle such as utilitarianism to determine the means of aggregation, without considering individual agents' preferences about ethical principles. Ethical principles are broader philosophies relevant to the group as a whole [14, 30]. For example, the principle of utilitarianism prioritises total gain for the group, while the personal value of traditionalism prioritises actions that reflect customs that an agent holds. Utilitarianism reflects an agent's belief about how a group or society should operate, whereas traditionalism focuses on how individuals within that society should act.

In comparison to value-aggregation, other methods of preference aggregation consider a single ethical principle or general goal. Some methods consider scenarios [12, 28, 32], where a group of decision makers bring about consensus through optimising for trust, or minimising disagreement. Rank aggregation methods [5] consider a MAS where each agent holds a ranking of results, and the principle used in aggregation is a set goal of the designer (e.g. reducing spam). This delegation can lead to challenges, as the designer may struggle to identify the most representative ethical principle or goal for a given situation, potentially introducing bias into multi-agent decision-making.

*Example 1.* Difference in principle preferences. Alice and Charlie are citizens in Decision City. There are two decisions that need to be made. First, the allocation of healthcare resources in a local hospital, and second, the location of a new station on a metro network.

Both Alice and Charlie hold preferences about personal values in these contexts. We model two personal values: hedonism and traditionalism. Charlie is more hedonistic, and Alice is more traditionalist. Charlie wants the station to be near him, and the hospital to allocate resources to his needs. Both citizens' decision logic are affected by the actions that are taken, i.e. if the station is built near Charlie, his value of hedonism will be promoted. Alice also wants the station to be built near her; Her value of traditionalism is not affected by the station decision, but her hedonism preference (although smaller than Charlie's preference) is still promoted.

Alice and Charlie hold preferences over ethical principles; Alice is an egalitarian, and Charlie a utilitarian. Alice strongly believes that resources should be assigned to ensure all patients are treated as equal in the local hospital, and would prefer if the metro stations were arranged to allow access for all the citizens in the city. If the metro stations are only built in the highest population centres but away from Charlie's house, Charlie's personal values are not promoted but his principle preferences are. Alice's principle preferences are demoted, as the smaller districts in the city have not been taken into consideration.

For the scenario illustrated in Example 1, previous methods of aggregation [11, 16, 24] consider Alice's and Charlie's personal values, but not their principle preferences. Incorporating principle preferences benefits the final decision by optimising for a solution that is more representative of the citizens complete views.

To the best of our knowledge no existing method of value aggregation captures an agent's own preferences of ethical principles; Current methods that utilise ethical principles leave the choice of principle(s) to the system designer.

Methods that consider multiple principles [11, 24] approach the problem by considering several principles at the same time rather than finding a single consensus principle like HCVA, and do not test for effectiveness of delegation of principles to agents.

*Contributions* Addressing key gaps in current value aggregation methods [11, 16, 24], we introduce a comprehensive approach, named Hierarchical Consensus Value Aggregation (HCVA). HCVA generates a single consensus ethical principle to aggregate agents' value systems. HCVA is representative of agents by integrating agents' preferences over both personal values and ethical principles. HCVA has applications in decision support, as well as autonomous systems where a group decision is required to be made that considers all agents simultaneously.

Using the European Social Survey (ESS) [6], we utilise EU citizens values and views on hypothetical EU policies to demonstrate the effectiveness of HCVA compared to a number of baselines. Baselines include aggregation following a single ethical principle [16], and aggregation according to a combination of many ethical principles [24]. Alongside ESS data, synthetic principle data is used to test HCVA's robustness across various principle preference sets. By establishing HCVA as a method motivated to be as adaptable to as many different ethical principles and values as possible, we envision that HCVA strengthens representivity for social choice ethics by considering the diversity of stakeholder preferences over ethical principles [23].

*Novelty* HCVA's novelty lies in moving away from existing methods that rely on an ethical principle being chosen by a central authority [11, 16, 24]. By incorporating each agent's preferences over principles, we reduce the biases that can arise from an authority's subjective choices. HCVA not only improves the fairness of the decision-making process but also provides a nuanced understanding of how different ethical perspectives influence group decisions.

*Organisation* Section 2 reviews previous related works. Section 3 presents our method, formalising value systems, value aggregation, and consensus generation. To evaluate our method we describe our experiment setup in Section 4 and analyse the results of said experiments in Section 5.1 and Section 5.2. We discuss these results in Section 5.3. Finally, Section 6 discusses conclusions and future work.

## 2 Preliminaries and Related Works

Works on value aggregation and preference aggregation that considers ethical principles are relevant, fitting into literature around value alignment.

### 2.1 Value Aggregation

Value Alignment [7, 15, 17] in group decision-making is the method of ensuring that a decision maker acts in a way that is reflective of human values. Coding normative ethical principles [30, 31] into decision-making tools is one method of ensuring a decision taken is value aligned. In social choice ethics, understanding how ethics should be defined in a particular context [4, 17] is important for fostering trust between the stakeholders in the system and the system itself.

Following the example situation set out in Section 1, we consider example cases to illustrate the problem of value aggregation.

*Example 2.* Different Ethical Principles depending on Context. Bob, another citizen, believes that decisions made in a local hospital should be more egalitarian, but decisions made for the metro should be more utilitarian. He holds preferences over ethical principles that change depending on the context.

Without considering Bob's complete preferences over various contexts, current methods of aggregation may not be representative of Bob's complete views.

*Example 3.* Personal Values not Reflecting Ethical Principles. Charlie, a steadfast utilitarian, may not support every action that a utilitarian decision-making system takes.

A decision-making system will aggregate personal values and make judgements according to the principle it aggregates on. Therefore, a central authority can abstract to state "Utilitarians prefer to take actions $a, b, c$ over $x, y, z$". However, such a method may not accurately represent all agents. By incorporating Charlie's personal values, he can express dissatisfaction with certain policies without abandoning his utilitarianism.

*Example 4.* Returning False Ethical Principles. Charlie may prefer that the city maximises the total happiness of agents (the ethical principle of utilitarianism). Under the Rawlsian veil of ignorance [20], Charlie does not know whether his principle preference will result in a group decision that maximises his own reward.

If his personal values are in the majority, then utilitarianism would benefit him, but egalitarianism would seek to minimise the contribution of a majority viewpoint, as a result of treating all agents as equal.[1]

### 2.2 Value Aggregation in Multi-Agent Systems

To resolve differences in preferences to make a fair decision for a group, Robinson [23] discusses a conceptual "AI decider", and distinguishes between moral solutions, compromise solutions and epistemic solutions. Unlike a *true moral theory approach* where a single moral theory is applied to the problem, previous work on value aggregation [11, 16, 24] has considered compromise solutions. These combine ethical principles by representing them as being a certain distance away from one another. Similarly, a combination approach [13] finds a solution that lies in-between principles.

[1] Note that when making this decision we aren't asking agents to rank the actions that could be taken, but asking about their preferences of values over every other value. We then ask their opinions on actions. If Charlie strongly supports an action, we say Charlie's values are promoted if that action is taken.

In a policymaking context, Serramia et al. [26] examine value alignment in participatory budgeting scenarios. A value alignment score is found that accounts for preferences, the relationships of those preferences to values, and the multiple proposals in the solution. Values are not general, such as the Schwartz system [25], and may lead to disagreements among agents regarding which policies promote or demote certain values.

Lera-Leri et al. [15, 16] utilise $L_p$ regression for aggregation, which we consider in this paper. $L_p$ regression is a method of computing linear regression according to a distance measure set by a value $p$. If $p = 1$ agents' values will be aggregated according to the $L_1$ norm. The authors consider different distance measures to be analogous to ethical principles following previous literature by González Pachon and Roméro [11]. Alternatively, Salas-Molina et al. [24] utilise $L_p$ regression in a "multi-$l_p$-norm approximation problem, aimed at minimising multiple $p$-metric distance functions" [24]. Given a list of $p$'s, their method can aggregate considering all values of $p$ simultaneously. Salas-Molina et al. succeed at creating a consensus that is balanced, taking into account each ethical principle fairly. In contrast, HCVA aggregates using a single value of $p$ that is derived from eliciting agents' principle preferences.

HCVA is a balanced solution using a *reflective equilibrium approach* to value-aligned group decision-making, achieving "an overarching and coherent moral view that incorporates as many of the ethical judgements of its decision subjects as possible" [23]. HCVA extends previous value aggregation approaches, considering the output of a reflective equilibrium approach on agent outcomes by incorporating agents' own ethical principles into value aggregation; The selection of an ethical principle to use in aggregation is therefore analogous to the selection of a voting system in an election.

Table 1 summarises related works in collective, value aligned group decision-making.

**Table 1.** Summary of current methods, considering the eliciting of data from agents and the use of principles in group decision-making. The limited number of current methods is due to the scope of this work only considering aggregation that handles ethical or moral preferences. Other preference based aggregation methods that do not consider ethical or moral preferences are not considered.

| | Elicits Agent's | | Considers Principle(s) | |
|---|---|---|---|---|
| | Values | Principles | One | Many |
| Serramia et al. [26] | ✓ | ✗ | ✗ | ✗ |
| González Pachon and Roméro [11] | ✓ | ✗ | ✓ | ✓ |
| Lera-Leri et al. [16] | ✓ | ✗ | ✓ | ✗ |
| Salas-Molina et al. [24] | ✓ | ✗ | ✓ | ✓ |
| HCVA (This Work) | ✓ | ✓ | ✓ | ✓ |

## 3 Hierarchical Consensus Value Aggregation

Here we provide an overview, formalisation and technical presentation of HCVA.

### 3.1 Brief Overview of HCVA

In HCVA, we consider the difference between ethical principles, personal values and actions to model agents accurately. In literature, ethical principles are applied as methods of changing how to aggregate values [16, 24, 11]; In our case, we use ethical principles as $L$-norm distance measures, aggregating values using $L_p$ regression. Values and actions on the other hand are represented as agents' preferences. Note that in HCVA, values are represented as pairwise preferences between each value and every other value. Actions are represented as the effect an action has on each value.

HCVA uses agents' *principle preferences* to define a single consensus ethical principle to use in the aggregation of agents' values and actions. We define two opposing ethical principles; Utilitarianism as $p = 0$ (the $L_0$-norm) and egalitarianism as $p = \infty$ (the $L_\infty$-norm). A consensus ethical principle can be a a combination of utilitarianism and egalitarianism, e.g. $p = 2.3$, which is handled by $L_p$ regression. We compute many preference aggregations using a range of different ethical principles. We find the average output of all preference aggregations to return a consensus ethical principle. Other strategies of finding an unbiased consensus ethical principle are discussed in section 6 as future work.

From this consensus ethical principle, we aggregate agents' values and actions to create a *consensus value system*—a value system that is representative of all agents views. Using the consensus value system we can make decisions on the actions considered. A pipeline of HCVA using the schematic defined below is shown in Figure 1.

### 3.2 Schematic of HCVA

We formally define a schema for elements required in computation. We draw from current literature [16, 24] in our definitions.

**Definition 1.** *Environment.* $E = \langle C, Ag, V, Pri \rangle$ *is a tuple that includes a set of contexts* $C$ *and a set of agents* $Ag$. $E$ *also contains a set of personal values* $V$ *and a set of ethical principles* $Pri$ *that can be reasoned about in the system. Values [25] and ethical principles are general, relevant across all contexts.*

**Definition 2.** *Context.* $c \in C$ *is the state of affairs that influence an agent's decision-making process and imposes constraints as to what actions that can be taken within the context.* $c$ *is a tuple* $c = \langle A_c, Ag_c \rangle$ *that contains the set of actions,* $A_c \subset A$*, that can be taken, and the set of agents* $Ag_c \subset Ag$*, that are considered in the context.*

**Definition 3.** *Agent.* $ag_i \in Ag$ *is a stakeholder in the environment, represented as a tuple* $a_i = \langle PVS_i, PriP_i \rangle$. *An agent is comprised of a personal value system* $PVS_i$ *and a set of principle preferences,* $PriP_i$. *The* $PVS_i$ *defines the agent's preferences about values and actions, whilst* $PriP_i$ *defines the agent's preferences about ethical principles.*

**Definition 4.** *Personal Value System.* $PVS_i = \langle VA_i, P_i \rangle$ *is a tuple belonging to an agent* $ag_i$, *representing their personal values.* $VA_i$ *contains the action judgements for each action given a particular value, representing the effect of an action on the agent's values.* $P_i$ *contains the agent's pairwise preferences over values.*

**Definition 5.** *Personal Value.* $v \in V$ *is a motivating factor for an agent, influencing their behaviour and decision making, where* $V = \{v_1, \ldots, v_x\}$ *is the set of all values, and* $x$ *is the number of values. Agents represent their preferences about values as a set of personal value preferences.*

**Definition 6.** *Personal Value Preferences.* $P_i$ *contains an agent* $ag_i$*'s set of pairwise value preferences* $P_i \in [0, 1]^{x \times x}$ *where* $x$ *is the number of values.*

$P_i$ *defines the importance of each value to agent* $ag_i$ *in a particular context* $c$. *For every value, there is a pairwise comparison with every other value* $P[a, b]$, *where a value preference of* $0.5$ *indicates indifference,* $0$ *indicates complete preference of value* $b$ *to* $a$, *and* $1$ *indicates complete preference of value* $a$ *to* $b$.

**Definition 7.** *Ethical Principle.* $pri \in Pri$ *is a principle that describes how the group should act. A principle such as utilitarianism is a type of normative principle that we map to a the* $L_1$ *distance measure, maximising total utility above all else. Egalitarianism on the other hand we map to the* $L_\infty$ *distance measure, maximising the welfare of the worst off agent. In this work, we only consider these two ethical principles and combinations of them. For example,* $L_{1.5}$ *is not completely utilitarian, but leans toward utilitarianism more than egalitarianism.*

**Definition 8.** *Principle Preferences.* $PriP_i$ *contains an agent* $ag_i$*'s set of pairwise principle preferences* $PriP_i \in [0, 1]^{y \times y}$ *where* $y$ *is the number of principles. The representation uses the same logic as pairwise value preferences.*

**Definition 9.** *Action.* $a \in A$ *is an action to be decided on that an agent reasons about in a context. The set of actions is defined as* $A = \{a_1, \ldots, a_z\}$ *where* $z$ *is the number of actions.*

**Definition 10.** *Action judgement.* $a_v^i : VA[-1, 1]^{x \times z}$. *Evaluates an agent* $ag_i$*'s promotion or demotion of every value* $v \in V$ *if an action* $a \in A$ *is taken. An action judgement of* $+1$ *indicates maximum promotion, and* $-1$ *indicates maximum demotion.* $0$ *is indifference.*

Figure 1 demonstrates how these elements are used in HCVA. HCVA uses the set of principle preferences to find a single ethical principle to aggregate the set of PVS's.

*Example 5.* Single Agent Formalism Example. In Decision City, Alice ($Al$) is asked her views on a number of contexts $c \in C$. Alice holds personal values of "traditionalism" and "hedonism," defined as $V = \{Tr,\ He\}$, and principles of "utilitarianism" and "egalitarianism", defined as $Pri = \{util,\ egal\}$. This forms the Environment $E = \langle C, Ag, V, Pri \rangle$, where $Ag = [Al]$.

In a context $c$, Alice is asked her views on a hedonistic policy. The action of deciding whether to support the policy is defined as $A = \{policy\}$. This forms the context $c = \langle V, A, Ag \rangle$. Alice prefers to act traditionalistic when considering this context, represented as pairwise value preferences $P_{Al}[Tr, He] = 0.8$, $P_{Al}[He, Tr] = 0.2$. Alice's action judgement matrix $VA^{Al}$ would contain two action judgements: $a^{Al}_{Tr}(policy)$ and $a^{Al}_{He}(policy)$, describing how her values are effected by taking the action. For Alice, $a^{Al}_{Tr}(policy) = -0.9$, as her value of traditionalism is negatively impacted by the policy, while $a^{Al}_{He}(policy) = 1$, reflecting how the policy is hedonistic (in Alice's view). This forms Alice's $\text{PVS}_{Al} = \langle VA_{Al}, P_{Al} \rangle$.

Alice is an egalitarian, and her principle preferences in this context $c$ are reflective, such that $PriP_{Al}[egal, util] = 0.9$, $PriP_{Al}[util, egal] = 0.1$. If more citizens of Decision City were to make the decision with Alice, these principle preferences indicate how Alice prefers the decision to be made. This forms Alice's agent $Al = \langle PVS_{Al}, PriP_{Al} \rangle$

Given that Alice strongly prefers traditionalism over hedonism in this context, and traditionalism is demoted by taking the action, not taking the policy is representative of Alice's values.

### 3.3 Aggregation using HCVA

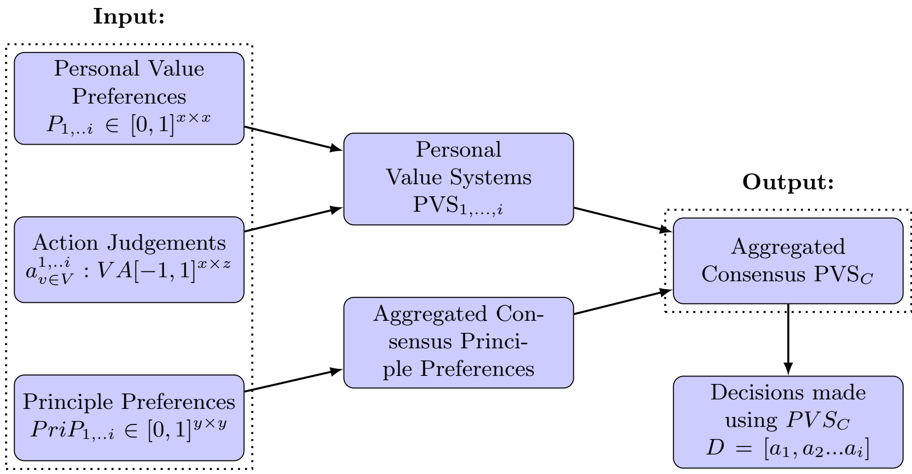
Here we formalise the aggregation of principle preferences and PVS's. The output of HCVA is consistent with prior literature [16, 11], outputting a consensus value system with a similar structure to an agent. This agent-representative format returned can be used to reason about decisions, as seen in Table 3.

*Aggregation of Principle Preferences* We first find the aggregation of the principle preferences $PriP_{1,..i} \in [0,1]^{y \times y}$ to find a consensus set of principle preferences representative of all agents. This consensus preference data can map to a singular ethical principle, which we use to aggregate the agents' PVSs.

A distance function is minimised to achieve fair aggregation. A generalised distance function $U^a_p$ is employed, following work by [10, 11, 16]. The principle preference matrix is represented as a $y$-dimensional vector, where $y$ is the number of principles in the vector. $U^y_p$ is general, where $y$ represents the data that we are aggregating at a specific time. $U^y_p$ can be used for pairwise preferences and action judgements by substituting $y$ for preference or action judgement data.

$$U^y_p = \left[ \sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \cdots \sum_{j_y=1}^{J_y} w_i |T_i[j_1, ..., j_y] - T_S[j_1, ..., j_y]|^p \right]^{\frac{1}{p}} \tag{1}$$

**Fig. 1.** Pipeline of HCVA. The PVS's of each agent is the set of their personal value preferences and action judgements. An agent's principle preferences are separated from their PVS. Principle aggregation is completed to find a consensus ethical principle. The agents' PVSs are aggregated twice using this consensus ethical principle (once for the action judgements, once for the preferences). While we use a simple formula to make decisions from the output of HCVA, we only evaluate against the consensus value system $\text{PVS}_C$.




Here, $n$ is the number of agents and $w_i$ is the weight vector for agent $ag_i$. $T_i[j_1, ..., j_y]$ represents the preference values $[j_1, ..., j_n]$ for the $i$-th agent for the set of features being considered; These can be principle preferences, value preferences or the action judgements for an agent $i$. $p \in [1, ..., \infty]$ corresponds to a p-metric distance function value [11]. Different values of $p$ in this equation map to different social choice functions. $p = 1$ maps to a utilitarian solution, $p = \infty$ maps to an egalitarian solution. We use $l_p$ regression [1] to solve $T_s = arg\ min\ U_p^{(T)}$, which is equivalent to minimising $x$ in $||Ax - b||_p$. A $p$ value of $p \in \{1, 2, \infty\}$ can be solved linearly [11], but for the sake of presentation, we solve all values of $p$ non-linearly such that we can include all potential values of $p$ in one general function $U_p^y$ [16]. This is solved using the iteratively re-weighted least squares (IRLS) algorithm. [2]

*Finding a Single p Value* The aim of aggregating principle preferences is to create a single consensus that minimises bias toward any principle. We minimise bias by aggregating agents' principle preferences repeatedly. We aggregate principle preferences using a range of $p$'s from 1.0 onwards in 0.1 intervals, stopping when the difference between $p_i$ and $p_{i+0.1}$ is less than a value of $\epsilon = 0.005$, defined arbitrarily; We find this value of $\epsilon$ produces good results. Using the list of con-

[2] We use Adil et al.'s [1] Python implementations of the IRLS algorithm.

sensuses returned by each aggregation we calculate the mean of the preference values, selecting the closest consensus to the mean (to 1 d.p.).

*Aggregation of the Personal Value Systems* Aggregation of the set of PVSs' involves computing a value system that represents the PVSs' of all agents. This is a two-step procedure, seen in Algorithm 1.

---

**Algorithm 1** Aggregation of the set of PVSs'

---

**Require:** For each agent $ag_i$: preference matrix $P_i \in \mathbb{R}^{x \times x}$ and action judgment matrix $VA_i \in \mathbb{R}^{x \times z}$
**Require:** $p \leftarrow$ Consensus from principle preference aggregation
1: $T_C^{(P)}$ ▷ Initialise empty consensus ($C$) for preferences
2: $T_C^{(VA)}$ ▷ Initialise empty consensus ($C$) for action judgements
3: $U_p^y =$ Equation 1 ▷ Initialise general aggregation function [16]
4: $T_C^{(P)} = arg\ min\ U_p^P$ ▷ Aggregate the Preferences
5: $T_C^{(VA)} = arg\ min\ U_p^{VA}$ ▷ Aggregate the Action Judgements
**return** $T_C^{(P)}, T_C^{(VA)}$

---

Figure 1 presents HCVA in full. Briefly,

(1) Retrieve context-specific data for agents' personal value preferences, action judgements and principle preferences. This makes up a set of PVS's and a set of principle preferences relevant to the context.
(2) Solve $U_p^y$ on the set of principle preferences using $l_p$ regression on different $p$ values, incrementing in 0.1 steps until the difference between $U_p^y - U_{p+0.1}^y < \epsilon$. Find the mean of these consensuses to return a single unbiased $p$ value to be used when aggregating all agents' PVSs.
(3) Solve $U_p^P$, $U_p^{VA}$ using $l_p$ regression with $p$ being the value found in the previous step to aggregate all agents' PVSs. This can then be considered the consensus PVS on which decisions can be made by the central authority.

## 4 Experimental Setup

We first explain the data used in experiments. Then, we explain baselines and metrics used to evaluate HCVA's performance.

### 4.1 European Social Survey

The European Social Survey (ESS) [6] is an annual attitudes survey completed by participating countries in Europe. Each country has a minimum sample size of 1,500 participants[3]. We employ questions regarding welfare attitudes and values from ESS 2016 in our experiments.

[3] Minimum number of participants is smaller for countries with a population of less than two million

We consider two questions that best map to the Schwartz values [25] of "traditionalism" and "hedonism". To describe these values, we use the questions "How important is it to follow traditions and customs," and "How important is it to have a good time"[4]. Only considering two values allows us to demonstrate the effect of each value on the consensus. Using these values, we consider the following action "Are you against, or in favour of an EU wide social benefit scheme?" We treat this question as one action, where a negative result would be against a benefit scheme, and a positive result would be in support. Table 2 shows a snippet of this data; Rudimentary analysis of the dataset shows a weak positive correlation between a country being more traditionalist and supporting the action.

We represent the ethical principles of utilitarianism and egalitarianism using the question "For a fair society, differences in standards of living should be small". As we consider utilitarianism and egalitarianism as opposites in HCVA's logic, we match the question to egalitarianism (treating people as equal as possible) and treat disagreement as utilitarianism[5]. This data serves as a real-world example and whilst is not a purpose-built mapping to these principles, does demonstrate the application of HCVA in real-world group decision-making.

We make two assumptions in our use of the ESS data. We assume that every respondent answers questions under the Rawlsian Veil of Ignorance [20], meaning that their answers are not influenced by their characteristics. We do this to avoid accounting for gaming of the system. An example of this could be an agent that realises that they are in the minority. Rather than provide accurate principle preferences, they may falsely describe themselves as more egalitarian to bolster their own personal values. We also do not consider weighting preferences in our experiments and all preferences are treated equally for simplicity.

### 4.2 Experiment Setup

Survey responses are grouped by country, and average data is used to create a set of representative PVS's and principle preferences. A snippet of this data is shown in Table 2.

*Baselines* We compare HCVA against four different $L_p$ regression based baselines. (1) $L_p$ regression where $p = 1$ corresponding to maximum utilitarianism (Labelled "Util"). (2) $L_p$ regression where $p = 10$ corresponding to maximum egalitarianism (Labelled "Egal"). As $p$ increases, the difference between $p$ and $p+0.1$ reduces. We treat $p = 10 \approx p = \infty$ for computation purposes. (3) A *transition point* $p$ between the utilitarian and egalitarian consensuses is calculated, labelled as "T" in results. This follows code from Lera-Leri et al. [16]. Many consensuses are computed with values of $p$ increasing in 0.1 increments, starting at $p = 1$. For each consensus the distance between the current $p$, the Util consensus ($p = 1$) and Egal consensus ($p = \infty$) is found. When the distances swap

[4] These questions are reworded slightly for conciseness

[5] When agents' represent entire countries, indifference responses are disregarded.

**Table 2.** Value preferences and action judgments for a sample of agents where values $V = \{Tr, He\}$ and the singular action $a = \{scheme\}$, making up the PVS for each country. Values selected have weak positive correlation between being traditionalist and supporting the action. Countries tend to be more egalitarian. Principle preference data is in the form of total responses.

| | Agent PVS | | | | Decision | Principle Pref. | |
|---|---|---|---|---|---|---|---|
| | $P[Tr, He]$ | $P[He, Tr]$ | $a_{Tr}(scheme)$ | $a_{He}(scheme)$ | Support | Egal | Util |
| AT | 0.35 | 0.65 | 0.05 | 0.11 | 0.09 | 1 394 | 188 |
| BE | 0.36 | 0.64 | −0.06 | −0.05 | −0.06 | 1 166 | 268 |
| CH | 0.54 | 0.46 | 0.15 | 0.26 | 0.20 | 939 | 221 |
| CZ | 0.28 | 0.72 | 0.02 | 0.03 | 0.03 | 893 | 683 |
| DE | 0.46 | 0.54 | 0 | 0.07 | 0.03 | 1 738 | 469 |

such that the $p$ value in question is closer to an Egal consensus, this is marked as the transition point, as seen in Figure 2. (4) An alternative method of value aggregation that uses $L_p$ regression and considers multiple $p$ values at the same time [24] (Labelled "SLM"). This is unlike HCVA, where a single representative value of $p$ is found from agent's preferences. The SLM baseline is passed a list of agents' $p$ values, which is found from the agent's principle preferences.

*Metrics* We compute three evaluation metrics. We focus on representativeness performance, but cover distribution of welfare to illustrate fairness differences between methods.

(1) To illustrate representativeness of agents we employ two voting systems. We utilise both proportionally representative and majority voting methods. Each agent votes on the consensus that is closest to their PVS, higher the better.
   - First Past The Post (FPTP) is used as a majority voting method for the best consensus out of all methods. Higher scores are better.
   - Borda count is used as a proportional representation method of voting for the best consensus. Higher scores are better.

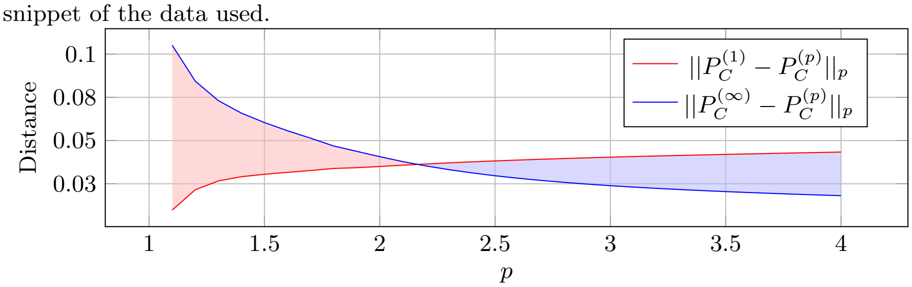
(2) To illustrate the distribution of welfare, we employ a residual calculation for each method as a fairness metric, evaluating the spread of satisfaction for each consensus. The absolute distance between every agent's PVS is measured against each consensus. Lower the better; meaning each agent is more aligned to the consensus produced by the method in question.

*Synthetic Principle Preferences* Alongside experiments using ESS data, we also generate synthetic principle preference data. We use synthetic data to test robustness by considering different edge cases not found in our real-world dataset. We consider how much an agents principle preferences *align* with their PVS. A principle will bring about consensuses that are in support of some values and against others. An informal example case can be seen in Example 3. Synthetic principle preferences support or oppose consensuses that are aligned to an agents personal values. We generate eight different principle preference sets as follows:

– (1) Upper and (2) Lower Quartile: Principle preferences are aligned to the agents PVSs', within the upper/lower quartile of strength.
– (3) Extreme Egal and (4) Util: All agents are egalitarian/utilitarian to some extent, independent to their PVS
– (5,6) General Support and (7,8) Opposition (75% and 50%): Principles support/oppose agents preferences respectively, within the top 75/50% of consensuses, such that a decision close to indifference is not counted.

We also compute the same experiments while treating individual respondents as agents to verify that the collation of individual responses by country is not biasing results in any way. This data can be found in the appendix A.

**Fig. 2.** Distance between a consensus computed using $L_p$ regression (x axis), and a fully utilitarian (Red, $L_p, p = 1$) and fully egalitarian consensus (Blue, $L_p, p = 10 \approx \infty$). The transition point is 2.2. As the value of $p$ tends to $\infty$, the difference in consensus generated between $p$'s reduces. For this reason, we treat $p = 10 \approx \infty$. Table 2 shows a snippet of the data used.




Notes that metrics are calculated only on each agents' PVSs. That is, their value preferences and their action judgements but not their principle preferences. Otherwise, HCVA performs best in all metrics.

## 5 Results

Here we present our results, separated by ESS and synthetic data.

### 5.1 Experiments with ESS Data

Consensuses computed by HCVA and baselines are found in Table 3. Table 4 and Figure 3 show voting results for each consensus.

*Key ESS Data Results.* (1) Egal has the worst residual results out of all strategies, corresponding to low welfare for agents. (2) HCVA is voted the most representative method by agents per Borda count and FPTP voting systems. This is closely followed by T. Both Util and Egal consensuses produce poor results.

**Table 3.** The aggregated consensus PVSs for ESS data [6] where $Tr$ is Traditionalist, $He$ is Hedonist, and *scheme* corresponds to the action. We calculate support for the decision using the following formula: for an action $a_i$ and agent $ag$: $score_i = (P_{ag}[v_1, v_2] \cdot a^{ag}_{v_1}(i)) + (P_{ag}[v_2, v_1] \cdot a^{ag}_{v_2}(i))$.

| Attributes | | Consensus PVS Elements | | | | Decision |
|---|---|---|---|---|---|---|
| Method | $P$-Value | $P[Tr, He]$ | $P[He, Tr]$ | $a_{Tr}(scheme)$ | $a_{He}(scheme)$ | Support |
| Egal | 10 | 0.37 | 0.63 | −0.05 | 0.01 | −0.01 |
| Util | 1 | 0.35 | 0.65 | 0 | −0.03 | −0.02 |
| T | 2.20 | 0.35 | 0.65 | −0.03 | 0.01 | −0.02 |
| SLM | N/A | 0.37 | 0.63 | −0.05 | 0.01 | −0.01 |
| HCVA | 2.50 | 0.36 | 0.64 | −0.03 | 0.01 | 0 |

(3) The SLM baseline performs the best in the residual baseline, but poorly in representative metrics. (4) As expected, HCVA, SLM, and T all outperform a Util or Egal consensus generally.

**Table 4.** Results for ESS survey data by method, (higher is better). Both a proportional representation and a majority voting system show that agents' overall vote HCVA the best consensus over all other baselines.

| | Egal | Util | T | SLM | HCVA |
|---|---|---|---|---|---|
| Borda | 18 | 27 | 70 | 39 | **76** |
| FPTP | 1 | 4 | 7 | 1 | **10** |

*ESS Data Results.* In both voting systems, HCVA produces a consensus that is more representative in the majority of cases. HCVA performs better than T because it is defined by the principle preferences of each agent. These preferences broadly map to agents' PVS's. T on the other hand is defined by the two extremes of the consensus, which is unrelated to agents' true intentions. SLM does not perform as well in voting metrics, which may be the result of HCVA and T taking the majority of votes where Util and Egal's extreme principles are not best, as SLM consistently ranks third in Borda scores. However, we did not expect SLM to underperform against T, as it still has the benefit of being a representative method. Like HCVA, SLM is defined by principle preferences of agents.

When comparing residuals (Figure 4), the Egal consensus has the largest residuals, and thus the worst welfare amongst agents. HCVA has a smaller interquartile range than T, although its median is higher, reflecting its $p$ value being more egalitarian than the transition point. SLM and HCVA balance attempting to receive the lowest variance between agents, whilst still outperforming other baselines on total utility for all agents in voting metrics. T does this also, but falls short of SLM and HCVA in terms of variance.

**Fig. 3.** Distribution of ranks received in Borda count voting (higher is better). Blue represents the highest ranking, and red the lowest. The ranking that an agent submits corresponds to the distance between an agent's PVS and the consensus. HCVA performs the best.

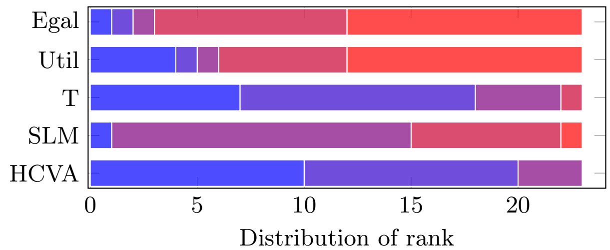
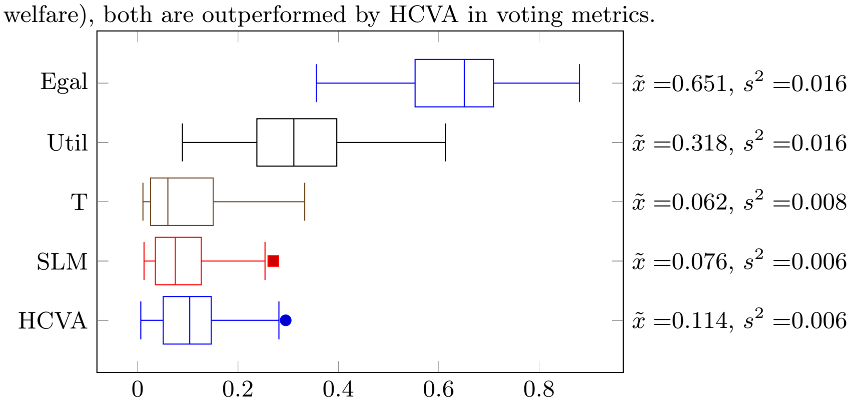



**Fig. 4.** Boxplots representing residuals for each agent for each baseline, with ESS Data. The residuals are calculated as the normalised distances between each agent's PVS and the PVS of the consensus. Right hand side shows median and variance values for each method rounded to three d.p. Whilst SLM and T have a better median value, and SLM has a closer interquartile range than HCVA (indicating a more egalitarian distribution of welfare), both are outperformed by HCVA in voting metrics.



### 5.2 Experiments with Synthetic Principle Preferences

Table 5 summarises the results from the synthetic principle preferences. We define how we generate synthetic data in section 4.2.

*Key Synthetic Principle Preference Results* (1) In scenarios where an agent's principle preferences are not aligned with its personal value preferences, T outperforms HCVA. However, when the ethical principle used strongly supports an agent's personal value preferences, HCVA demonstrates superior performance. (2) While HCVA is outperformed by SLM when the ethical principle used opposes agents' personal value preferences, it consistently ranks second in all other scenarios, showcasing its robustness. (3) HCVA yields mixed residual results, reflecting its adaptability across various voting methods.

*Synthetic Principle Preference Results* T excels in synthetic data where principles conflict with agents' PVSs. However, HCVA outperforms T when principles are strongly aligned with personal values, highlighting its effectiveness in these contexts. Although Egal and Util consensuses show slight improvements, they still fall short of both the T and HCVA overall in all metrics.

In experiments with misaligned principles, HCVA and SLM do not benefit from principle preferences that accurately reflect agents' PVSs. This means the resulting consensus is less representative with lower welfare than T, defined only by the agents' PVSs. Nevertheless, HCVA remains a strong contender, consistently ranking as the "best of the rest" in scenarios except when principle preferences oppose personal value preferences. SLM performs similarly to HCVA but falls away in the FPTP metric. Overall, HCVA maintains a competitive edge in most scenarios and proving its utility in real-world applications of group decision-making.

**Table 5.** FPTP and Borda scores for each strategy (Higher is better) in the format <FPTP, Borda>. T performs best in all but the Upper Quartile case. HCVA performs "best of the rest".

| | Egal | Util | T | SLM | HCVA |
|---|---|---|---|---|---|
| Lower Quartile | 2, 21 | 3, 25 | **15, 82** | 1, 45 | 2, 57 |
| Upper Quartile | 1, 18 | 4, 27 | 7, 70 | 1, 39 | **10, 76** |
| Egal | 1, 18 | 4, 27 | **11, 75** | 1, 47 | 6, 63 |
| Util | 2, 25 | 1, 24 | **15, 79** | 1, 50 | 4, 52 |
| Support 75% | 1, 18 | 3, 30 | **12, 76** | 1, 49 | 6, 57 |
| Opposition 75% | 1, 18 | 3, 31 | **15, 79** | 1, 53 | 3, 49 |
| Support 50% | 2, 25 | 1, 23 | **15, 79** | 1, 51 | 4, 52 |
| Opposition 50% | 1, 18 | 3, 31 | **15, 79** | 1, 53 | 3, 49 |

*Additional Results* In the appendix, Figure A7 presents the residual box-plots for synthetic society data. Looking at individual values in Figure A5 and Tables A3 and A4, the character of the data is replicated at an individual level. Each respondent in the data considers each strategy as an individual, and rankings are collated. This is true when also considering synthetic data, as seen in Tables A1, A2, A3, A4 and Figure A5.

### 5.3 Discussion

Incorporating agents' principle preferences allows for a more nuanced and representative consensus. As seen in the examples in section 2, whilst ethical principles and PVS's are separate, there is a relationship between the two. This is useful in the cases where there are gaps in data collected, as a broader ethical principle can help interpret an agent's PVS. However, this method may not accurately represent all agents. Therefore, value-aligned AI in group decision-making should consider both principles and personal values, as decisions are complex and multifaceted.

Considering an agent's principles may increase or decrease fairness depending upon the outcome. If a decision to an agent is fair, the decision doesn't infringe on an agent's core values, principles or capabilities [3, 18, 19, 22]. We define fairness mathematically by looking at the distribution of welfare, using the residual calculation found in Section 5. We show that HCVA is fairer than baselines using real world data. A method that combines different ethical principles together seems to create the fairest solution in our cases, with varying success depending on the context and strategy utilised. However, it is unlikely that a single ethical principle is the most fair consensus.

## 6 Conclusion and Future Work

We address a key limitation of value aggregation to enable representative value-aligned decisions among pluralistic agents. Evaluation with ESS data [6] and synthetic societies demonstrate the robustness and fairness of HCVA in decision-making processes, considering agents' personal values as well as principle preferences. Consensuses that consider combining ethical principles seem to perform better overall, however HCVA improves representativeness by considering the disconnect between personal values and principle preferences as shown in Section 2.

### 6.1 Future work

**Calibration:** By using a range of aggregations on agents' principle preferences, we intend to minimise the weighting of one distance measure over any other to give an unbiased aggregation. This is not a completely unbiased method and instead returns a diluted aggregation that still contains some biases. As the value of $\epsilon$ increases in HCVA, the consensus $p$ will be weighted more to lower values of $p$. Other methods of aggregation that are representative and unbiased, such as one proposed by Bana et al. [2] would create a more unbiased approach to finding a consensus $p$ value.

**Types of Values in Reflective Equilibrium Methods:** Aggregating values is further complicated by the fact that some values are *protected*, whereby an agent's values are considered non-negotiable. In literature, [4, 8, 21], values are not general [25] but instead are contextual e.g. "it should be categorically prohibited to deny anyone the right to drive however they like" [4]. These values "arise from deontological rules,..., concerning actions rather than consequences" [21]. Baron [3] continues to define protected values as non-negotiable social norms. By using basic human values like those described by Schwartz [25], HCVA can separate non-negotiable values (such as all cases that the European Convention on Human Rights (ECHR) [27] prohibits) from the aggregation, avoiding an impossible aggregation with non-negotiable values. This is a form of safeguarding that has to potential to mitigate the pitfalls found in reflective equilibrium approaches like HCVA.

*Reproducibility* Code can be found at https://github.com/JosephTrevorrow/Hierarchical-Consensus-Value-Aggregation. Appendix A provides further analysis.

**Acknowledgments.** JT is supported by the UK Research and Innovation (UKRI) Centre for Doctoral Training in Interactive Artificial Intelligence Award (EP/S022937/1). NA thanks UKRI EPSRC Grant No. EP/Y028392/1: AI for Collective Intelligence (AI4CI) for the support. We thank Emanuele Ratti, Enrico Liscio, Jessica Woodgate and Roger Lera-Leri for their advice and suggestions.

## References

1. Adil, D., Peng, R., Sachdeva, S.: Fast, provably convergent IRLS algorithm for p-norm linear regression. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 14189–14200. Curran Associates Inc., Vancouver (2019). https://doi.org/10.5555/3454287.3455558
2. Bana, G., Jamroga, W., Naccache, D., Ryan, P.Y.A.: Convergence Voting: From Pairwise Comparisons to Consensus (Mar 2021). https://doi.org/10.48550/arXiv.2102.01995, http://arxiv.org/abs/2102.01995, arXiv:2102.01995 [cs]
3. Baron, J.: Protected Values and Other Types of Values. Analyse & Kritik **39**(1), 85–100 (May 2017). https://doi.org/10.1515/auk-2017-0005, https://www.degruyter.com/document/doi/10.1515/auk-2017-0005/html, publisher: De Gruyter Oldenbourg
4. Baum, S.D.: Social choice ethics in artificial intelligence. AI & SOCIETY **35**(1), 165–176 (Mar 2020). https://doi.org/10.1007/s00146-017-0760-1, https://doi.org/10.1007/s00146-017-0760-1
5. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the Web. In: Proceedings of the 10th International Conference on World Wide Web. pp. 613–622. WWW '01, Association for Computing Machinery, Hong Kong, Hong Kong (Apr 2001). https://doi.org/10.1145/371920.372165, https://dl.acm.org/doi/10.1145/371920.372165

6. European Social Survey ERIC (ESS ERIC): European social survey (ESS), round 8 - 2016 (2017). https://doi.org/10.21338/NSD-ESS8-2016
7. Gabriel, I.: Artificial intelligence, values, and alignment. Minds and Machines **30**(3), 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2
8. Ginges, J., Atran, S., Medin, D., Shikaki, K.: Sacred bounds on rational resolution of violent political conflict. Proceedings of the National Academy of Sciences **104**(18), 7357–7360 (May 2007). https://doi.org/10.1073/pnas.0701768104, https://www.pnas.org/doi/abs/10.1073/pnas.0701768104
9. González-Pachón, J., Romero, C.: Distance-based consensus methods: a goal programming approach. Omega **27**(3), 341–347 (1999). https://doi.org/10.1016/S0305-0483(98)00052-8, https://linkinghub.elsevier.com/retrieve/pii/S0305048398000528
10. González-Pachón, J., Romero, C.: Aggregation of ordinal and cardinal preferences: A framework based on distance functions. Journal of Multi-Criteria Decision Analysis **15**, 79–85 (05 2008). https://doi.org/10.1002/mcda.426
11. González-Pachón, J., Romero, C.: Bentham, Marx and Rawls ethical principles: In search for a compromise. Omega **62**, 47–51 (2016). https://doi.org/10.1016/j.omega.2015.08.008, https://www.sciencedirect.com/science/article/pii/S0305048315001632
12. Hassani, H., Razavi-Far, R., Saif, M., Herrera-Viedma, E.: Consensus-Based Decision Support Model and Fusion Architecture for Dynamic Decision Making. Information Sciences **597**, 86–104 (Jun 2022). https://doi.org/10.1016/j.ins.2022.03.040, https://www.sciencedirect.com/science/article/pii/S0020025522002523
13. Leben, D.: A Rawlsian algorithm for autonomous vehicles. Ethics and Information Technology **19**(2), 107–115 (Jun 2017)
14. Leben, D.: Normative principles for evaluating fairness in machine learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. p. 86–92. AIES '20, Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3375627.3375808
15. Lera-Leri, R., Bistaffa, F., Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.: Towards pluralistic value alignment: Aggregating value systems through lp-regression. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. pp. 780–788. International Foundation for Autonomous Agents and Multiagent Systems, Virtual Event, New Zealand (2022), https://dl.acm.org/doi/abs/10.5555/3535850.3535938
16. Lera-Leri, R.X., Liscio, E., Bistaffa, F., Jonker, C.M., Lopez-Sanchez, M., Murukannaiah, P.K., Rodriguez-Aguilar, J.A., Salas-Molina, F.: Aggregating value systems for decision support. Knowledge-Based Systems **287**, 111453 (2024). https://doi.org/https://doi.org/10.1016/j.knosys.2024.111453, https://www.sciencedirect.com/science/article/pii/S0950705124000881
17. Martin, D.: Who Should Decide How Machines Make Morally Laden Decisions? Science and Engineering Ethics **23**(4), 951–967 (Aug 2017). https://doi.org/10.1007/s11948-016-9833-7, https://doi.org/10.1007/s11948-016-9833-7
18. Nussbaum, M.: Capabilities and human rights. In: De Greiff, P., Cronin, C.P. (eds.) Global Justice and Transnational Politics, pp. 273–300. The MIT Press (2002), https://doi.org/10.7551/mitpress/3302.003.0007
19. Ratti, E., Graves, M.: A capability approach to ai ethics. American Philosophical Quarterly **62**(1), 1–16 (01 2025). https://doi.org/10.5406/21521123.62.1.01, https://doi.org/10.5406/21521123.62.1.01

20. Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge, Massachusetts, 2nd edition edn. (1999)
21. Ritov, I., Baron, J.: Protected Values and Omission Bias. Organizational Behavior and Human Decision Processes **79**(2), 79–94 (Aug 1999). https://doi.org/10.1006/obhd.1999.2839, https://www.sciencedirect.com/science/article/pii/S074959789992839X
22. Robeyns, I.: The Capability Approach: a theoretical survey. Journal of Human Development **6**(1), 93–117 (Mar 2005). https://doi.org/10.1080/146498805200034266
23. Robinson, P.: Moral disagreement and artificial intelligence. AI & SOCIETY **39**(5), 2425–2438 (Oct 2024). https://doi.org/10.1007/s00146-023-01697-y, https://doi.org/10.1007/s00146-023-01697-y
24. Salas-Molina, F., Bistaffa, F., Rodríguez-Aguilar, J.A.: A general approach for computing a consensus in group decision making that integrates multiple ethical principles. Socio-Economic Planning Sciences **89**, 101694 (2023). https://doi.org/https://doi.org/10.1016/j.seps.2023.101694, https://www.sciencedirect.com/science/article/pii/S0038012123002069
25. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online Readings in Psychology and Culture **2**(1), 3–20 (Dec 2012). https://doi.org/10.9707/2307-0919.1116
26. Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Moretti, S.: Value alignment in participatory budgeting. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. p. 1692–1700. AAMAS '24, International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand (2024)
27. European convention on human rights (1950), https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=005
28. Tsiporkova, E., Boeva, V.: Multi-step ranking of alternatives in a multi-criteria and multi-expert decision making environment. Information Sciences **176**(18), 2673–2697 (Sep 2006). https://doi.org/10.1016/j.ins.2005.11.010, https://www.sciencedirect.com/science/article/pii/S0020025505003270
29. Witesman, E.M., Walters, L.C.: Modeling public decision preferences using context-specific value hierarchies. The American Review of Public Administration **45**(1), 86–105 (2014). https://doi.org/10.1177/0275074014536603, https://journals.sagepub.com/doi/10.1177/0275074014536603
30. Woodgate, J., Ajmeri, N.: Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. ACM Computing Survey **56**(11), 289:1–289:37 (Jul 2024). https://doi.org/10.1145/3672394, https://dl.acm.org/doi/10.1145/3672394
31. Woodgate, J., Ajmeri, N.: Combining normative ethics principles to learn prosocial behaviour. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. pp. 2789–2791. AAMAS '25, International Foundation for Autonomous Agents and Multiagent Systems (2025), https://dl.acm.org/doi/abs/10.5555/3709347.3744013
32. Xu, Z., Cai, X.: Group consensus algorithms based on preference relations. Information Sciences **181**(1), 150–162 (Jan 2011). https://doi.org/10.1016/j.ins.2010.08.002, https://www.sciencedirect.com/science/article/pii/S002002551000366X

# A Appendix

Here we include tables for individual respondents (whereby each agent in our environment represents a single survey response, rather than representing all responses in a country) to show the consistency of trends between the two representations. We also include figures containing boxplots that show the distribution of residuals for HCVA in synthetic societies.

**Table A1.** Borda scores for each strategy for individual respondents in Austria using synthetic principle data described in Section 4 (Higher is better). This data is shown to demonstrate that the same trends appear on individual survey respondents as well as grouped by country.

| | HCVA | T | Util | Egal | SLM |
|---|---|---|---|---|---|
| Bottom 25 | 4271 | 6609 | 1701 | 2097 | 3902 |
| Top 25 | 6474 | 5529 | 1634 | 1831 | 3112 |
| Exreme Egal | 5743 | 5648 | 1791 | 1723 | 3675 |
| Extreme Util | 3440 | 6567 | 1542 | 2464 | 4567 |
| General Support 75 | 5128 | 6082 | 1929 | 1676 | 3765 |
| General Opposition 75 | 4324 | 6244 | 2072 | 1649 | 4291 |
| General Support 50 | 3440 | 6567 | 1542 | 2464 | 4567 |
| General Opposition 50 | 4324 | 6244 | 2072 | 1649 | 4291 |
| ESS Data | 6474 | 5529 | 1634 | 1831 | 3112 |

**Table A2.** FPTP results for each strategy by individual respondents in Austria using synthetic data described in Section 4 (Higher is better).

| | HCVA | T | Util | Egal | SLM |
|---|---|---|---|---|---|
| Bottom 25 | 186 | 1224 | 159 | 255 | 34 |
| Top 25 | 1051 | 501 | 74 | 232 | 0 |
| Exreme Egal | 988 | 612 | 134 | 124 | 0 |
| Extreme Util | 234 | 1335 | 0 | 255 | 34 |
| General Support 75 | 656 | 990 | 135 | 77 | 0 |
| General Opposition 75 | 509 | 1138 | 161 | 50 | 0 |
| General Support 50 | 234 | 1335 | 0 | 255 | 34 |
| General Opposition 50 | 509 | 1138 | 161 | 50 | 0 |
| ESS Data | 1051 | 501 | 74 | 232 | 0 |

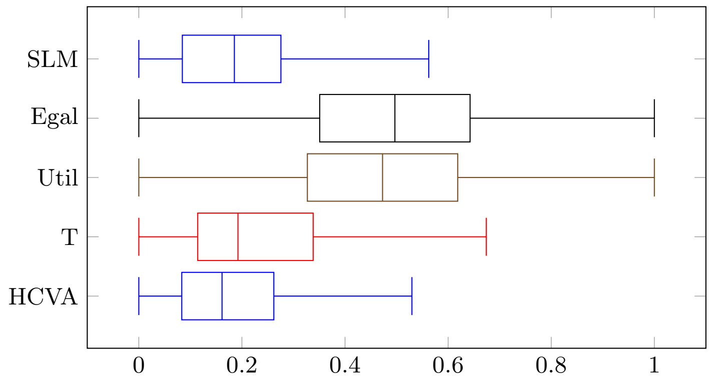



**Fig. A5.** Boxplots representing residuals for individual respondents for each strategy, using ESS data. The residuals are calculated as the normalised distances between each agents PVS and the value system of the consensus, described in Section 4. Whilst SLM and T have a better mean, and SLM has closer interquartile range than HCVA, both are outperformed by HCVA in voting metrics.

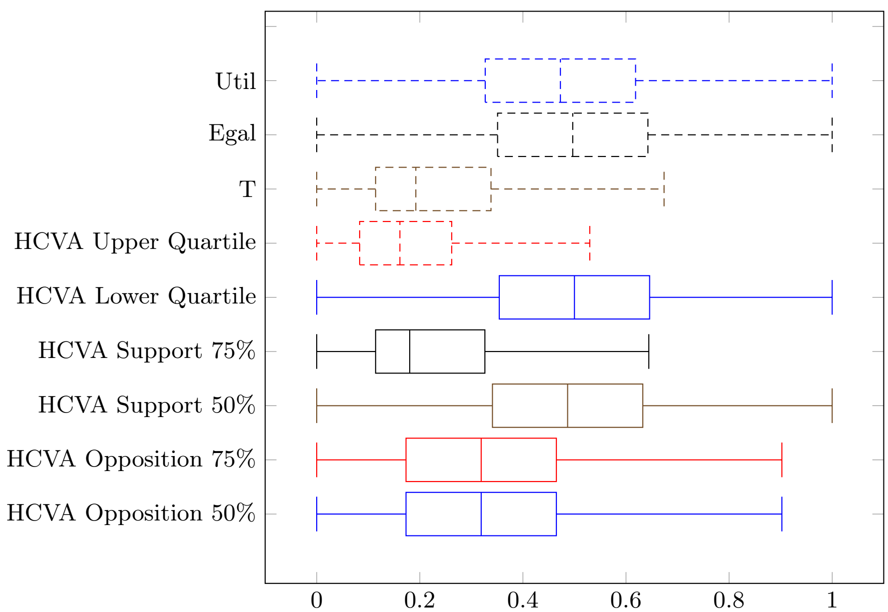



**Fig. A6.** Boxplots representing residuals for individual survey respondents for consensuses generated by HCVA for each type of synthetic principle data described in Section 4. SLM has been omitted for clarity. T, Egal and Util Consensuses have been included as baselines. The spread of residuals computed by HCVA show poor performance in comparison to T, but generally performs "best of the rest" compared to baselines. Data using individual survey respondents is presented to show that trends on country data is replicated in individual survey data.

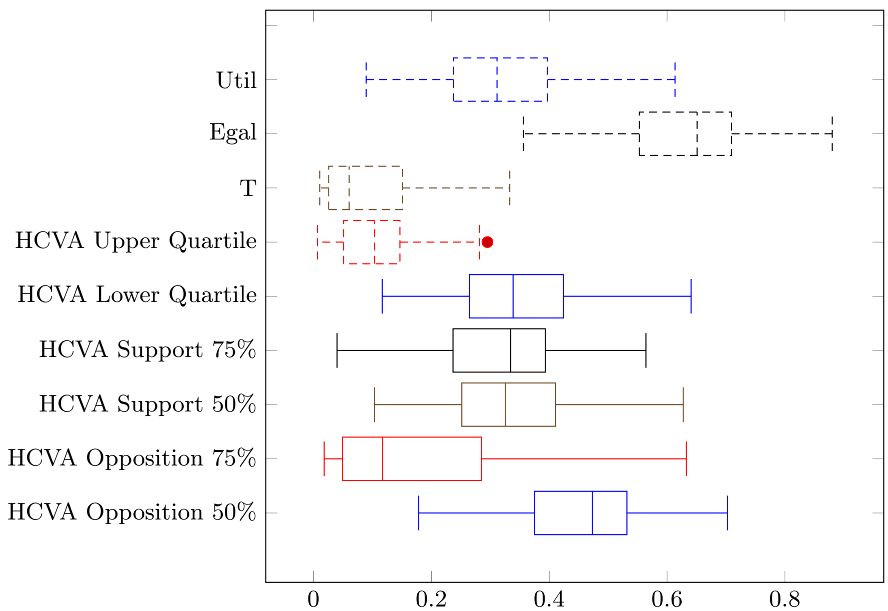



**Fig. A7.** Boxplots representing residuals for each country for consensuses generated by HCVA using synthetic principle data described in Section 4. SLM has been omitted for clarity. T, Egal and Util consensuses have been included as baselines. The spread of residuals computed by HCVA show poor performance in comparison to T, but generally performs "best of the rest" compared to baselines.

**Table A3.** Borda scores for each strategy for every individual response in ESS Data using synthetic principle data described in Section 4 (higher is better). This demonstrates the same trends as seen in the case where we group respondents by country.

| | HCVA | T | Util | Egal | SLM |
|---|---|---|---|---|---|
| Bottom 25 | 96994 | 143895 | 38929 | 40401 | 79171 |
| Top 25 | 138785 | 121944 | 38931 | 34928 | 64802 |
| Exreme Egal | 121348 | 124311 | 42472 | 33218 | 78041 |
| Extreme Util | 77392 | 143313 | 35691 | 47561 | 95433 |
| General Support 75 | 108101 | 133668 | 45098 | 32486 | 80037 |
| General Opposition 75 | 91049 | 136908 | 48600 | 32202 | 90631 |
| General Support 50 | 77392 | 143313 | 35691 | 47561 | 95433 |
| General Opposition 50 | 91049 | 136908 | 48600 | 32202 | 90631 |
| ESS Data | 138785 | 121944 | 38931 | 34928 | 64802 |

**Table A4.** FPTP results for every individual respondent of the survey using synthetic principle data described in Section 4 (higher is better).

| | HCVA | T | Util | Egal | SLM |
|---|---|---|---|---|---|
| Bottom 25 | 3845 | 27560 | 3238 | 4752 | 544 |
| Top 25 | 23004 | 11119 | 2164 | 3652 | 0 |
| Exreme Egal | 21376 | 13343 | 3278 | 1942 | 0 |
| Extreme Util | 5041 | 29602 | 0 | 4752 | 544 |
| General Support 75 | 13722 | 21729 | 3278 | 1210 | 0 |
| General Opposition 75 | 10514 | 24779 | 3720 | 926 | 0 |
| General Support 50 | 5041 | 29602 | 0 | 4752 | 544 |
| General Opposition 50 | 10514 | 24779 | 3720 | 926 | 0 |
| ESS Data | 23004 | 11119 | 2164 | 3652 | 0 |