



CITY UNIVERSITY
LONDON

MSc Artificial Intelligence Master's Thesis.

Víctor Abia Alonso

student ID: 230061642

March 7, 2025

Contents

1	Introduction	1
2	Relevant context.	2
2.1	Pluralistic value alignment	2
2.2	Discussion on the state of the art	3
2.3	Preference aggregation.	5
3	Methodology	5
3.1	Problem Definition	5
3.1.1	Desirable Properties of the Aggregation Function	6
3.2	Methods	7
3.3	Assumptions and scope.	8
4	Results	9
4.1	Framework for the solution	9
4.1.1	Definition of a Value	9
4.1.2	The notion of context	9
4.1.3	Value System	9
4.1.4	Moral Utility of a Value System	10
4.2	Both aggregation methods	11
4.2.1	Approach to solving the problem: p -norm minimization.	11
4.2.2	Value-oriented aggregation (VOA)	11
4.2.3	Action-oriented aggregation (AOA)	12
4.3	Solving the optimization problem.	12
4.3.1	Matrix Representation of both methods.	12
4.3.2	Convexity Analysis	13
4.3.3	Numerical Solution	14
4.4	Aggregation function analysis: are VOA and AOA the same?	15
4.5	Analytical solution: conditions and special cases.	17
4.5.1	First Order Necessary Conditions	17
4.5.2	Mathematical Derivation of Consensus Weights for $p = 2$	19
4.5.3	Mathematical Derivation of Consensus Weights for $ V = 2$	21
4.5.4	Mathematical Derivation of Consensus Weights for $ H = 2$	22
4.6	Toy Example: Urbanism	23
4.7	European Value Study data	25
5	Discussion and Conclusion	26
5.1	Practical use cases: which value of p and which aggregation method?	27
5.2	Future Work	27

1 Introduction

Morality is a fundamental concept of humankind which guides our decisions as individuals and society. The history of philosophy showcases different approaches to understand and reason about moral arguments and scenarios. Nowadays, the advancement of artificial intelligence (AI) and the increased autonomy and proactivity of agents, emphasize the need to get a deeper understanding of our moral values so that AI systems can be guided by the values of humanity. The literature has studied the problem of value alignment (Russell 2019), the problem of aligning AI behaviour and decisions with a given set of moral values and preferences over them. However, an important question still remains: *What moral values should the AI align with?* Thus, creating useful mathematical models of the moral values of people and society is still an open problem which is crucial for aligning AI and ensuring its safe development.

Given AI's decisions can affect vast groups of people, it is paramount that the values governing the behavior of AI represent all of them. Hence, this project looks into **value system aggregation**, the process of aggregating several individual value systems (i.e. the individual's values and preferences over them) into a common value system that can represent the group of people as a whole. In other words, value aggregation seeks to combine individual moral perspectives into a unified moral stance that can be used in AI applications representing the group as a whole (the so-called 'consensus value system'). This problem first requires of some mathematical framework to successfully capture the moral preferences of an individual, for which there are already some examples in the literature. But also, it requires reasonable mathematical functions to perform the aggregation so that it is grounded in useful principles. The literature has not provided any principled approach, hence this project aims at filling this gap. Since in most cases values are used to make action decisions, the presented approach follows a novel action-oriented approach to ensure that decisions made by the common aggregated values are as similar as possible to the action decisions made by the individuals' values.

The main contributions of this project are as follows:

- A **novel definition of value systems** based on weighted preferences, which allows for a more granular representation of individual moral value systems and is needed for translating value systems into utility functions over actions.
- A new aggregation approach called *Value-Oriented Aggregation* (VOA): This approach builds upon existing methods in the literature, focusing on aggregating individual preferences based on value hierarchies.
- The introduction of a **new principle for aggregation functions** called Value Utility Alignment Principle. This principle ensures that the aggregated values lead to decisions that minimize discrepancies with the decisions individuals would have made based on their personal value systems.
- Another novel aggregation approach called *Action-Oriented Aggregation* (AOA). This new approach is designed in line with the novel Value Utility Alignment principle, aiming to aggregate individual preferences in a way that directly considers the actions influenced by the value systems.
- A **rigorous mathematical analysis** of both aggregation methods, including mathematical proofs of convexity and the analytical solution to edge cases. The analysis also includes guidance on how to set the parameters for these methods.
- A **proof of concept** and practical illustration to demonstrate how these methods can be applied in real-world settings. This includes a comparative analysis, a toy example, and a real-world case study using survey data.

- The **Python code** to numerically solve both aggregation methods, allowing others to calculate value aggregation using the AOA or VOA methods.

Note however that while the code provides practical tools for applying the aggregation methods, the primary contribution of this work lies in its rigorous mathematical analysis, which advances the understanding of value aggregation and moral value systems in the literature.

The outcomes of this research are expected to have far-reaching implications, particularly in areas where moral value aggregation plays a pivotal role, such as artificial intelligence alignment and participatory politics. By providing novel methods for aggregating values, this project offers practical tools for decision-making in contexts like participatory budgeting, where the moral values of citizens guide societal choices (Serramia et al. 2024). These contributions are especially significant for future generations, given the growing importance of AI alignment in mitigating the worst-case risks associated with artificial intelligence. Furthermore, this work benefits academic fields like value systems research by advancing the state of the art through a fundamentally novel approach. Additionally, the mathematical proofs and results derived in this project can also benefit social choice theory—and other fields where similar frameworks to the one used here for value aggregation—to address complex decision-making problems.

The project is now divided into the following four sections. In Section 2, a literature review is done to lay out the most up-to-date research on value systems and value aggregations, as well as clarifying some key ideas for the reader. In Section 3, the aggregation problem is formally described, along with other key tools used to carry out the project. Section 4 includes all the aforementioned contributions in the same order. Finally, Section 5 considers the limitations of the methods, discusses its relevance, and suggests next steps.

2 Relevant context.

2.1 Pluralistic value alignment

The rapid advancement of AI has brought forth critical ethical challenges, one of which is the AI value alignment problem (Russell 2019, Sierra et al. 2019), i.e. ensuring that AI systems align with human moral values. The field of Trustworthy AI (Chatila et al. 2021, Commission 2019, European Union 2024) emphasizes the need for AI systems to reflect the values of the societies they serve. Osman et al. (Osman & d’Inverno 2024) argue about the necessity of computational models that represent and reason over human values, highlighting the need for formal value systems, that is a set of moral values and their preferences. These value systems enable AI to align with the diverse and sometimes conflicting values of human agents.

In fact, this idea that people’s ethical stance is composed of several key different values that reflect deeper human motivations has been analysed in several disciplines. In sociology, Schwartz’s theory of basic human values (Schwartz 2012) identifies ten universal values, such as benevolence, tradition, and security, which are shared across all societies, providing a framework for understanding social behaviors. In psychology, Moral Foundations Theory (Haidt 2012) proposes that human moral reasoning is built on six fundamental foundations, such as care, fairness, and loyalty, which vary in emphasis across cultures and individuals. In applied ethics (Beauchamp & Childress 2019) (Ross 1930), values are understood to be context-dependent, with different values providing distinct judgments on actions based on the circumstances and the moral principles relevant to the decision at hand. All of these views understand human decision-makers as value systems, thus going beyond the traditional view in economics that models human preferences as the utility provided by the actions to the agent. Note that here the terms agent, stakeholder, decision-maker, individual, person, and human are used interchangeably because they represent a value system as, in essence, we are only interested in the moral values they hold.

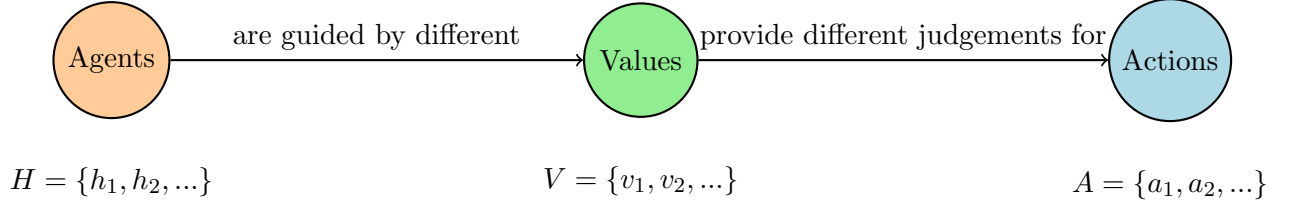


Figure 1: This figure shows that agents make decisions according to several values, which in turn judge actions differently. This is the general setup for pluralistic value alignment. In a given context C , we have a finite set of actions A , a finite set of values V which judge each of the actions and, for value aggregation, also a finite set of agents H (let’s say H for ”human”, but it is not limited to humans).

Understanding the values of a stakeholder can be useful by itself in circumstances that involve solely one individual like a personal assistant (e.g. Alexa), to understand your privacy preferences and adjusting the data collected accordingly (Serramia et al. 2023). However, the impact of AI is not usually bounded to one individual, hence to ensure it behaves as expected ”by the society” or ”by the group of affected individuals” it is important that their common values are considered. Value aggregation aims to resolve this plurality issue by providing the value system of a group from their individuals’ value systems. This representative value system can then serve to align AI with the will of the group. Despite the extensive research on preference aggregation and voting systems (see Brandt et al. (2016) for an overview) however due to the complex nature of values, usual preference aggregation functions are not useful for this task. In fact, there is only one proposed approach to tackle the issue of value system aggregation (Lera-Leri et al. 2022, 2024).

2.2 Discussion on the state of the art

For AI alignment it is usually assumed that we know which values to align the AI with. This line of work of capturing human values is called **value inference** and is usually been visualized in three steps (Liscio et al. 2023): identification, estimation and aggregation.

In different contexts, different values are important. For example, transparency is important in the context of governance but not in the context of medical records where privacy is the relevant value. Hence, the first step, called **value identification**, considers a specific context C with a finite set of specific actions A , and aims at identifying the relevant values V for that context. Researchers have tackled this problem using a hybrid approach based on Natural Language Processing (NLP) and human annotators to analyze context-specific text (Liscio et al. 2021). Through vector embeddings and cosine similarity, key values were identified based on opinions from a large corpus of context-specific text, facilitating value identification in context. For example, in the context of city planning, related books and essays can be fed so that the values of accessibility, green areas, fairness, individual freedom and public services could be detected as relevant.

After identifying the relevant values, the next step is value system estimation, in other words, to learn the value system of an individual. A value system is composed of a set of values (those identified in the previous step) and the individual’s preferences over them, therefore this step is focused on gauging how important each value is for the individual. This involves having some sort of preference structure over values like weighted preferences, simply an ordering, or an even simpler pairwise comparison between the values. Current methods for **value preference estimation** consist in presenting the individual with situations where they have to make a decision, and then analyse participants’ choices and the motivations they provide resolving inconsistencies between the two to derive accurate value preferences (Liscio et al. 2024). Additionally, values

are usually considered to have different meanings for each individual (Serramia et al. 2023b) so estimating the value understandings (a.k.a value interpretations or value judgements) is also needed to complete the value systems estimation. This is thought to be required because even for the same value in the same context there could be radically different interpretations about how that value judges an action. For example, people in the US and Europe have polar opposite understandings of the value of security with regard to carrying firearms.

The final step in capturing a group’s moral preferences, and thus the values an AI should align with, is **value aggregation**. Lera-Leri et al. (Lera-Leri et al. 2022, 2024) are the first to address this explicitly, proposing a two-step optimization method that aggregates individual value preferences and understandings. This results in a consensus value system, structurally identical to individual systems, but composed of aggregated understandings and preferences. While their approach demonstrates the feasibility of aggregating value systems, it lacks a guiding principle, raising concerns about whether such aggregations are suitable for decision-making.

Aggregating different value understandings for the same value aims to give an intermediate value interpretation to capture its meaning for the group. However, the aggregated value preferences are produced from each individual’s preferences without accounting for differences in the understanding of the value. Thus, in Lera-Leri et al.’s approach both aggregations (the one over understandings and the one over preferences) are effectively performed independently while they actually are closely related. Previously, we noted the stark difference of understanding of the value of security with regard to firearms, aggregating these opposing understandings would result in an intermediate interpretation of a value that agents wouldn’t have considered important for this action. Given the polar opposite understanding of security, the aggregated preference over these different understandings would be not meaningful to guide behavior (i.e. to decide if carrying a weapon is acceptable or not). In essence, the **previous work for value aggregation hasn’t proposed any principled approach** to the problem, which is what this project aims to do.

Having different value understandings leads to values that are less meaningful to the agents. To avoid this, I propose a framework where **all value understandings are shared**. The existing methodology for value inference, as explained above, already provides value name-tags, several related keywords, and an overall goal for each value (Liscio et al. 2021). Thus, to perform meaningful value aggregation, this global characterization—with keywords and an overall goal—should guide the shared understanding of a value within a given context. These values are inferred for a specific context and are useful only within that context. To prevent distortion caused by allowing agents to interpret values differently and further distorting them during value aggregation, defining a shared understanding for each value in the given context ensures a robust, structured meaning. This ensures that the value retains its meaning during aggregation. For example, there shouldn’t be multiple interpretations of the value $v = \text{security}$ but rather different values like $v_1 = \text{private security}$ and $v_2 = \text{collective security}$, representing distinct interpretations. Therefore, the goal of value inference should be to identify all meaningful principles guiding decision-making, rather than broad values with multiple interpretations.

So far, research in value aligned AI has been divided into two interdependent research blocks. The value system inference research is focused on obtaining value systems that can be used by AI (as previously discussed in Section 2.2) (Liscio et al. 2023, 2021, Siebert et al. 2022, Lera-Leri et al. 2022). Then, another body of research uses those value systems to align AI behaviour to values using, for example, value-aligned norms (Montes & Sierra 2021, Serramia et al. 2018), decision-making approaches (Serramia et al. 2023c), approaches to resolve ethical dilemmas (Soto et al. 2022), and many more. However, the first block of research has not considered how design decisions impact the second block of research. Indeed, when designing value inference approaches, their impact on AI value-aligned behaviour has been disregarded so far. In contrast,

this project presents a value system aggregation approach which considers how the aggregated value system favors or disapproves actions relevant in the context of application.

2.3 Preference aggregation.

Still, these shared value understandings should guide the aggregation and the different judgements of each value to each action should be considered. For that, I propose a novel structure of value preferences method not considered before using **weighted preferences** of the values, guiding the explicit calculation of a utility of an action using value systems by an agent; successfully connecting Agents to Actions in the pluralistic value systems context of Figure 1.

Preference aggregation has been extensively studied in social choice theory (Brandt et al. 2016), with much of the foundational work relying on the use of ordered rankings to represent individual preferences—i.e., preferring option X over Y, over Z—without reflecting the strength of these preferences. This preference framework also dominates in the current literature on value alignment (Serramia et al. 2020) (Siebert et al. 2022). However, ordered rankings alone fail to capture the intensity with which individuals prioritise certain values. This limitation can lead to significant misalignments between an agent’s preferences over potential actions and the preferences inferred by its value system, due to an underrepresentation of the magnitude of value preferences. As a result, modeling agents with value systems as guiding criteria must go beyond simple value rankings, particularly when the aim is to evaluate actions effectively.

Weighted preferences provide a solution by allowing us to quantify the relative importance of each value, thus indicating how much each value contributes to decision-making, instead of assuming equal weight across values. This concept aligns with the **moral parliament model of ethics** (Newberry & Ord 2021), where values “vote” on actions with varying degrees of influence depending on their assigned weights. Introducing weighted preferences into value systems adds the necessary structure to evaluate actions more precisely, enabling a better alignment of aggregated preferences with the actual decision-making process of agents.

3 Methodology

This section outlines the methodological framework for the value aggregation process. We aim to formally define the aggregation of individual value systems into a consensus value system while preserving their inherent structure. Value systems are defined in a given context C , and the aggregation process ensures that the consensus system represents the collective values in that same context. As it was explained in Section 2.2, several structures of value systems have been proposed in the literature. Hence, in this section we formalize the aggregation for a general value system structure and in the next section we provide a solution with a suitable structure. Finally, the technical methods to solve the problem, as well as the assumptions, are presented.

3.1 Problem Definition

Before delving into the definition of the value system aggregation problem, we shall first provide a general definition of a value system:

Definition 3.1 (Value System). *A value system $\mathcal{V} = \langle V, \succeq \rangle$ is a set of values V and a preference relation over these values \succeq .*

Definition 3.2 (Value Aggregation Function). *Let $H = \{h_1, h_2, \dots, h_{|H|}\}$ be a set of agents, each represented by an individual value system \mathcal{V}_k in a context C . If we note as \mathcal{VS}_C the set of all possible value systems in this context, then the value aggregation function F is defined as:*

$$F : \mathcal{VS}_C^n \rightarrow \mathcal{VS}_C$$

The function F can take as input any finite number n of individual value systems and returns a consensus value system $\mathcal{V}^S \in \mathcal{VS}_C$:

$$\mathcal{V}^S = F(\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_{|H|})$$

This definition conceptualizes the aggregation function as a mapping from a multidimensional set of value systems to a single consensus value system. To ensure that this function operates correctly and fairly, several additional properties can be imposed.

3.1.1 Desirable Properties of the Aggregation Function

We now define the following key properties for the aggregation function F :

Property 3.3 (Commutativity). *The aggregation function F is commutative if the order in which individual value systems are aggregated does not affect the final consensus. Formally, for any permutation function π over the agent indexes $\{1, \dots, |H|\}$:*

$$F(\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_{|H|}) = F(\mathcal{V}_{\pi(1)}, \mathcal{V}_{\pi(2)}, \dots, \mathcal{V}_{\pi(|H|)})$$

Note that the above property ensures anonymity i.e. that no agent has more influence based solely on its identity.

As previously mentioned, the goal of value system aggregation is to produce a consensus value system that can be used to make decisions regarding the actions AI should perform, ensuring alignment with everyone’s values. The following is a novel property that can serve as a principle to guide the value aggregation process, in contrast to previous value aggregation approaches Lera-Leri et al. (2022, 2024), which do not follow any particular principle. Before that, we first define formally utility, in order to get a sense of how value systems can relate to actions.

Definition 3.4 (Utility of a value system.). *Let A be a set of actions in a context C , and H the set of agents each with a corresponding value system in \mathcal{VS}_C . A value utility function $u : A \times \mathcal{VS} \rightarrow \mathbb{R}$ assigns a real-valued utility to each action based on how the action aligns with the agent’s value system and their preferences over these values.*

Property 3.5 (Value Utility Alignment Principle). *Let $d(\cdot, \cdot)$ be a distance function over two outputs of a value utility function $d : \mathbb{R}^2 \rightarrow \mathbb{R}$. Then the consensus value system $\mathcal{V}^S \in \mathcal{VS}_C$ is said to satisfy the value utility alignment principle if for a given set H of agents:*

$$\mathcal{V}^S = \arg \min_{\mathcal{V}} \sum_{k=1}^{|H|} \sum_{j=1}^{|A|} d(u(\mathcal{V}, a_j), u(\mathcal{V}_h, a_j))$$

The minimizing function F aims at minimising the difference between the utility of each action with regard to each individual value system and the aggregated one.

3.2 Methods

The commutativity and distance minimization properties ensure that the aggregation function F operates consistently, providing a mathematically sound approach for generating a consensus value system. With this foundation, the aggregation problem is formally defined. The goal of this project is to establish a reasonable structure for value systems, ensuring that the value aggregation process is meaningful. Two novel aggregation functions, Value-Oriented Aggregation (VOA) and Action-Oriented Aggregation (AOA), are introduced, each satisfying some or all of the aforementioned properties. These functions are mathematically analyzed, and a proof of concept demonstrating their application is provided. The main methodology to achieve this involves mathematical formalization and theoretical analysis of the proposed functions.

While the primary methodological approach for this project involves mathematical formulations and theoretical analysis of the value aggregation problem, additional tools and techniques have been employed to support the investigation and enhance the robustness of the results. These methods include further mathematical frameworks, optimization tools, and data analysis techniques, as outlined below:

- **Extended Mathematical Formulations:** Beyond defining the aggregation problem, mathematical structures are crucial for solving it. This involves expressing the aggregation processes in formal terms, enabling both qualitative and quantitative reasoning. For instance:
 - **Linear algebraic** reformulation of aggregation methods into matrix form enables computational approaches to solving the aggregation problem.
 - **Multivariate analysis** techniques can be employed to explore the minimization of distance functions, so that solutions meet the project’s theoretical requirements.
 - **Comparative analysis** of different aggregation methods through specific metrics allows evaluation of the magnitude of the differences between consensus outcomes.
- **Mathematical Proofs:** Formal proofs will be conducted to validate core theoretical results, including:
 - **Demonstrating the convexity** of both Value-Oriented Aggregation (VOA) and Action-Oriented Aggregation (AOA) methods according to their provided definition.
 - **Proving equivalence** between both aggregation methods under certain scenarios, which in turn allows to identify in which cases they yield relevantly different outcomes.
- **Optimization Techniques:** Numerical optimization will play a key role in calculating solutions to the aggregation problem, particularly for non-analytical cases. For this, Python libraries such as `numpy` and `scipy.optimize` become essential:
 - The consensus solution will be obtained by minimizing objective functions defined over the aggregation methods, employing the state-of-the-art **BFGS algorithm**.
 - **Constrained optimization** techniques are incorporated, ensuring that the solution remains feasible within the context of the problem.
- **Data Analysis:** A real-world application of the methods will be illustrated through survey data analysis:
 - **Data processing** of the survey responses will be performed using Python’s `pandas` library, with steps including data cleaning and transformations.
 - **Visualizations** of tables and graphs will allow for deeper insight into the aggregation methods solutions.

3.3 Assumptions and scope.

This project is based on several key assumptions that define the structure and scope of the value aggregation model. By formalizing individual moral value systems and representing them through structured weights, we aim to create a coherent framework for aggregating these values into a collective decision-making model. The assumptions detailed below ensure that the aggregation process is meaningful, consistent, and aligned with real-world decision-making processes, while also acknowledging the limitations of modeling complex human moral judgments through formalized, mathematical structures.

- **The consensus moral value system can be inferred from the individual ones.** The goal of this project is to quantitatively aggregate the moral preferences of individual agents in order to create a practical decision-making model that automates the process of deriving a consensus moral value system. This approach assumes that individual value systems are fixed and that, through mathematical aggregation, a collective moral framework can be constructed. While this method offers a structured, automated solution to consensus-building, it stands in contrast to other decision-making processes such as group deliberation and assembly-based approaches, where values and preferences can evolve through dialogue and mutual understanding. These alternative methods, often found in social movements and deliberative democracies, prioritize consent and open-ended conversation to reach consensus, allowing for dynamic shifts in individual preferences. In (Dryzek 2000), this trade-off is highlighted between the flexibility of the deliberative processes and the efficiency, clarity, and consistency of aggregation approaches like the one adopted in this project.
- **For a given context, there exists a finite set of values with a common interpretation, reflecting each person’s moral values.** We assume that all agents operate with the same interpretation of a given set of values in a specific context, allowing for a meaningful aggregation of preferences. This ensures that the values being compared are well-defined across individuals, enabling consensus. While this assumption might appear restrictive, it reflects real-world situations where people often share the same core values but differ in how they prioritize them. For example, individuals with differing political ideologies often prioritize the same core values in distinct ways, even if they share a common understanding of what those values represent (Graham et al. 2013). In cases where there is disagreement on the so-called understanding of a value, it may be that different values are being labeled similarly.
- **Human moral values, within a context, can be represented as weighted preferences that inform a utility function over actions.** This project assumes that human moral values, for a given context, can be expressed through a structured set of weights that assign relative importance to each value. The use of weights is crucial for translating individual value systems into actionable preferences over specific decisions, ensuring consistency in the aggregation process. While individuals may not explicitly reason in terms of weighted values, this approach captures the implicit prioritization that guides moral decision-making. By using weights, we avoid arbitrary functions that could introduce bias and instead create a framework where value systems align closely with the actual decisions made by individuals. This unexplored assumption is key to constructing a "moral utility function" that reflects both the individual and collective preferences over actions.

4 Results

This section encapsulates all the contributions of this work, including the proposed framework, the design of the solution methods, formal proofs, comparison with random baseline models, and a proof of concept. It considers a group of agents that, within a given context, operate based on a set of values, which in turn guide their decisions over a predefined set of actions.

4.1 Framework for the solution

The following subsections provide the formal definitions of values, context, value systems, and their associated utility functions, establishing the foundations for the aggregation methods discussed in subsequent sections.

4.1.1 Definition of a Value

In the framework of value pluralism, a *value* represents one of the fundamental moral principles or guidelines held by an agent. These values often conflict with one another in decision-making scenarios. In this work, following the practical ethics literature (Chisholm 1963), we define a value within a specific context C as a pair of judgment functions over actions. Specifically, each value v is characterized by two judgment functions (v^+, v^-) , which map actions to a numerical scale:

$$v^+, v^- : A \rightarrow [-1, 1]$$

Here, v^+ represents the positive evaluation of an action, while v^- represents the negative evaluation. These functions must satisfy the constraint that, for every action a ,

$$v^+(a) \cdot v^-(a) \leq 0$$

This ensures that an action positively evaluated by v^+ cannot simultaneously be positively evaluated by v^- , though it is not required that $v^-(a) = -v^+(a)$. This way judgements are constrained to be logically consistent within the applied ethics framework.

4.1.2 The notion of context

In ethical decision-making, values are always interpreted within the scope of a specific context. A context provides the background circumstances, including the relevant factors and actions that shape the moral discourse. More formally, a context C consists of a set of possible actions A and a set of values V , each with an associated interpretation (v^+, v^-) , that guide decision-making. We assume this interpretation is shared by all agents in order to allow a meaningful aggregation.

Here, we understand that what defines a context is the interpretation of the relevant values associated with it, which is a flexible definition that can be applied in practice to most situations. By asking "What are the relevant moral values to consider in this situation (where we have the set of available actions A)?" we can start to understand that situation as a context for value systems.

4.1.3 Value System

A value system represents a set of values held by an agent, along with the preferences that indicate the relative importance of each value. For this work, we adopt a structured view where the preferences among values are represented by weighted quantitative measures. In any given

context C , where there is a defined set of values V and a set of available actions A , the value system of the k -th agent is formalized as:

$$\mathcal{V}_k = \langle V, \mathbf{w}_k \rangle$$

Here, $V = \{v_1, \dots, v_{|V|}\}$ is the set of values, each of which is associated with a pair of judgment functions (v_i^+, v_i^-) , and $\mathbf{w}_k = \{w_1^k, \dots, w_{|V|}^k\}$ represents the weights assigned to each value by agent k , such that:

$$\sum_{i=1}^{|V|} w_i^k = 1$$

For visualization, a value system can be represented as a matrix of size $|V| \times 2|A|$, where each row corresponds to a value v_i and each column corresponds to the positive and negative evaluations of the available actions. With one additional column the weights w_i^k can be included, representing the preferences of the agent.

An example with 3 values and 2 actions is shown below:

$$M = \begin{pmatrix} v_1^+(a_1) & v_1^-(a_1) & v_1^+(a_2) & v_1^-(a_2) & w_1^k \\ v_2^+(a_1) & v_2^-(a_1) & v_2^+(a_2) & v_2^-(a_2) & w_2^k \\ v_3^+(a_1) & v_3^-(a_1) & v_3^+(a_2) & v_3^-(a_2) & w_3^k \end{pmatrix} = \begin{pmatrix} 0.8 & -0.1 & -0.7 & 0.2 & 0.50 \\ -0.5 & 0.3 & -0.6 & 0.9 & 0.25 \\ 0.3 & -0.2 & 0.4 & -0.5 & 0.25 \end{pmatrix}$$

In this representation, each row corresponds to a value so that the last column represent the relative importance of the value to the agent and rest represents the meaning of that value with respect to actions which is shared by all agents in that context. We will not represent value systems like this for the mathematical operations because its more useful to capture in 3 different matrices the positive judgements G^+ , the negative ones G^- and the weights of all agents W .

4.1.4 Moral Utility of a Value System

In this work, we define a clear notion of the preference for an action according to a value system. Previous research (Serramia et al. 2023b), proposed a transition from ordinal rankings to cardinal representations to quantify the relevance of values. While there is no universal method for converting ordinal preferences to cardinal ones, the inclusion of weighted preferences in our framework allows us to explicitly capture the importance of each value.

The global positive judgment function J_k^+ for the k -th agent is defined as the weighted sum of the positive judgments for each action:

$$J_k^+(a) = \sum_{i=1}^{|V|} w_i^k \cdot v_i^+(a), \quad J_k^-(a) = \sum_{i=1}^{|V|} w_i^k \cdot v_i^-(a)$$

where $J_k^+(a)$ and $J_k^-(a)$ represent the agent's overall positive and negative evaluations of action a , respectively. These functions satisfy $J_k^+, J_k^- \in [-1, 1]$, as $\sum_{i=1}^{|V|} w_i^k = 1$ and $v_i^+(a), v_i^-(a) \in [-1, 1]$.

These global judgment functions provide a way to translate an agent's preferences over values (via weights w_i^k) into preferences over actions (utilities). Specifically, $J_k^+(a)$ and $J_k^-(a)$ reflect the agent's utility for the action being performed or not, based on their value system.

Thus, a value system is described not only by the preferences among values—represented by the weights—but also by the agent's preferences over actions—captured by the global judgment functions. This dual representation suggests two possible approaches to aggregation: one focusing on aligning the value weights with the consensus and the other on aligning the global preferences.

4.2 Both aggregation methods

For both aggregations, we are aiming to get a consensus value system (with weighted preferences) that synthesizes the individual ones.

4.2.1 Approach to solving the problem: p -norm minimization.

In any aggregation process, it's essential to define a metric that quantifies how different the elements to be aggregated are. The p -norm (or l_p -norm) provides a flexible way to measure the distance between two vectors. Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the p -norm distance is defined as:

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

For $p = \infty$, this distance becomes the maximum absolute difference between the vector components:

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$$

The p -norm is especially useful because key aggregation metrics like the mean, median, and mid-range of a given sample arise as the values that minimize this distance function for different values of p . As shown in (González-Pachón & Romero 2016), when $p = 1$, the value that minimizes the absolute distances corresponds to the median; for $p = 2$, the value that minimizes the square distances corresponds to the mean; and for $p = \infty$, the value that minimizes the maximum distances corresponds to the mid-range. These three cases correspond to the utilitarian, the egalitarian and the fairness solutions respectively.

For example, considering the one-dimensional values 5, 9, and 10: minimizing the $p = 1$ distance results in 9 (the median), $p = 2$ gives 8 (the mean), and $p = \infty$ yields 7.5 (the mid-range). In the multidimensional case, these metrics are applied independently to each dimension of a vector.

In value aggregation for AI alignment, the p -norm allows us to measure the distance between individual value systems and a consensus system. Both aggregation methods presented—Value-Oriented Aggregation (VOA) and Action-Oriented Aggregation (AOA)—use the p -norm to minimize the some differences between agents' value systems and their consensus, providing flexibility based on the choice of p .

4.2.2 Value-oriented aggregation (VOA)

Value-oriented aggregation stems from the principle of achieving consensus based on similar value preferences. Specifically, it seeks to find the consensus value system whose weights are closest to the individual agents' weights. This similarity is measured using the p -norm distance between the individual weights and the consensus weights. The method is formalized by the objective distance function U_p^V , which sums the p -norm distances between the value weights w_i^k of each agent k and some candidate weights $\mathbf{w}^{s'} \in \mathcal{VS}_C$. The consensus value system, denoted as \mathbf{w}^S , consists of the set of weights $\{w_i^S\}_{i=1}^N$ that minimize U_p^V .

$$U_p^V = \left[\sum_{k=1}^{|H|} \sum_{i=1}^{|V|} |w_i^k - w_i^{s'}|^p \right]^{1/p}, \quad \mathbf{w}^S = \arg \min_{\mathbf{w}^{s'}} U_p^V$$

This method focuses solely on the preferences between values (i.e., the weights) and does not consider the judgment functions associated with each value.

4.2.3 Action-oriented aggregation (AOA)

Similarly, action-oriented aggregation stems from the principle of achieving consensus based on similar utilities of actions (Property 3.5). Specifically, it seeks to find the consensus value system whose global judgement functions are similar to the individual agents' global judgement functions. This similarity is measured using the p -norm distance metric between the individual judgements and the consensus judgements. The method is formalized by the objective distance function U_p^A , which sums the p -norm distances between the judgement functions J_k^+ , J_k^- of each agent k and the judgement $J_{s'}^+$, $J_{s'}^-$ resulting of some given weights $\mathbf{w}_{s'}$. The consensus value system, denoted as \mathbf{w}^S , consists of the set of weights $\{w_i^S\}_{i=1}^N$ that minimize U_p^A .

$$U_p^A = \left[\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} |J_k^+(a_j) - J_{s'}^+(a_j)|^p + |J_k^-(a_j) - J_{s'}^-(a_j)|^p \right]^{1/p}, \mathbf{w}^S = \arg \min_{\mathbf{w}_{s'}} U_p^A$$

Unlike U_p^V , U_p^A is calculated iterating over every agent and every action and AOA thus considers the judgement of values in its objective distance function. Additionally, U_p^A is iterating over the values implicitly because all the judgement functions $J(a)$ are a weighted sum of the judgement of all the values, and are these optimal weights in specific \mathbf{w}^S which constitute the consensus value system. By construction, this aggregation method satisfies the Value Utility Alignment Principle (Property 3.5).

4.3 Solving the optimization problem.

Both VOA and AOA, translate the value aggregation problem into an optimization problem by defining an objective function which is minimized by the consensus weights. Thus, in order to look at what is the proposed consensus value system by both methods, we need to solve this minimization problem. Note that this is a constrained optimization problem as both methods find the optimal weights \mathbf{w}^S from the set of all feasible weight distributions $\mathbf{w}^{s'}$ where all weights need to be positive and add up to one.

In order to tackle this minimization problem, we first show the convexity of both objective functions U_p^V and U_p^A with respect to the weights of the consensus $\mathbf{w}^{s'}$, which are the ones being optimized. For this, we vectorise the problem considering a matricial version of it. Convexity is a sufficient condition of proving that solutions are unique, which makes both optimization problems well-posed. Once this is shown, numerical solutions are considered.

4.3.1 Matrix Representation of both methods.

To redefine U_p^V in a matricial form, we define the following vectors and matrices:

- \mathbf{W} : The weight matrix of size $|V| \times |H|$, where $|V|$ is the number of values and $|H|$ is the number of agents. Each column \mathbf{w}^k in \mathbf{W} represents the weights for agent k , and overall this matrix describes all the individual values preference (which then define their value system).
- \mathbf{w}^S : The consensus weight vector of size $|V| \times 1$, which we aim to optimize.
- $\mathbf{1}^T$: A row vector of ones of size $1 \times |H|$ used to facilitate broadcasting operations.

- $\Delta \mathbf{W}$: The difference matrix of size $|V| \times |H|$, which is defined as $\Delta \mathbf{W} = \mathbf{W} - \mathbf{w}^S \mathbf{1}^T$. This operation subtracts the consensus weight vector from each column of \mathbf{W} , resulting in the signed residuals matrix.
- $\|\cdot\|_p$: The ℓ_p -norm, $\|\mathbf{x}\|_p = (\sum_i |x_i|^p)^{1/p}$ to measure the deviations in the residuals matrix.

$$\mathbf{W} = \begin{bmatrix} w_1^1 & w_1^2 & \cdots & w_1^{|H|} \\ w_2^1 & w_2^2 & \cdots & w_2^{|H|} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|V|}^1 & w_{|V|}^2 & \cdots & w_{|V|}^{|H|} \end{bmatrix}, \quad \mathbf{w}^S = \begin{bmatrix} w_1^S \\ w_2^S \\ \vdots \\ w_{|V|}^S \end{bmatrix}, \quad \mathbf{1}^T = [1 \quad 1 \quad \cdots \quad 1]$$

The objective function U_p^W can then be written in matrix form as:

$$U_p^W = \left(\sum_{k=1}^{|H|} \sum_{i=1}^{|V|} |w_i^k - w_i^S|^p \right)^{1/p} = \left(\sum_{k=1}^{|H|} \sum_{i=1}^{|V|} |\Delta W_{i,k}|^p \right)^{1/p} = \|\Delta \mathbf{W}\|_p$$

Where $\Delta W_{i,k}$ represents the element in the i -th row and k -th column of the matrix ΔW .

To minimize U_p^A , we define the following vectors and matrices:

- \mathbf{W} , \mathbf{w}^S , $\mathbf{1}^T$, and $\Delta \mathbf{W}$: The same as before.
- \mathbf{G}^+ : The positive judgement matrix of size $|V| \times |A|$, where $|V|$ is the number of values and $|A|$ is the number of actions. Each element $v_i^+(a_j)$ represents the positive judgement for value v_i and action a_j . It is chosen to be the G matrix instead of the J matrix, not to confuse it with the global judgement function.
- \mathbf{G}^- : The negative judgement matrix of size $N \times M$, analogous to \mathbf{J}^+ but using the negative judgement $v_i^-(a_j)$.

$$\mathbf{G}^+ = \begin{bmatrix} v_1^+(a_1) & v_1^+(a_2) & \cdots & v_1^+(a_M) \\ v_2^+(a_1) & v_2^+(a_2) & \cdots & v_2^+(a_M) \\ \vdots & \vdots & \ddots & \vdots \\ v_{|V|}^+(a_1) & v_{|V|}^+(a_2) & \cdots & v_{|V|}^+(a_M) \end{bmatrix}, \quad \mathbf{G}^- = \begin{bmatrix} v_1^-(a_1) & v_1^-(a_2) & \cdots & v_1^-(a_M) \\ v_2^-(a_1) & v_2^-(a_2) & \cdots & v_2^-(a_M) \\ \vdots & \vdots & \ddots & \vdots \\ v_{|V|}^-(a_1) & v_{|V|}^-(a_2) & \cdots & v_{|V|}^-(a_M) \end{bmatrix}$$

The objective function U_p^A can then be written in matrix form as:

$$U_p^A = \left[\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} |(\mathbf{D}^+)_{k,j}|^p + |(\mathbf{D}^-)_{k,j}|^p \right]^{1/p}$$

Where:

$$\mathbf{D}^+ = \Delta \mathbf{W}^T \mathbf{G}^+ \quad \text{and} \quad \mathbf{D}^- = \Delta \mathbf{W}^T \mathbf{G}^-$$

4.3.2 Convexity Analysis

To demonstrate that U_p^V is convex and has a unique solution, we consider the following points:

- **Objective Function Convexity:**

- The ℓ_p -norm $\|\cdot\|_p$ is a convex function for $p \geq 1$.

- The residuals matrix $\Delta \mathbf{W} = \mathbf{W} - \mathbf{w}^{s'} \mathbf{1}^T$ is affine in $\mathbf{w}^{s'}$ because \mathbf{W} is constant and $\mathbf{w}^{s'}$ is the variable.
- The composition of a convex function (the ℓ_p -norm) with an affine function (the residuals) is convex.

- **Constraints Convexity:**

- The constraints $\sum_{i=1}^{|V|} w_i^{s'} = 1$ and $w_i^{s'} \geq 0$ define a simplex ($|V|$ -dimensional tetrahedron), which is a convex set.

- **Strict Convexity and Uniqueness:**

- For $p > 1$, the ℓ_p -norm is strictly convex.
- Therefore, the objective function is strictly convex, ensuring a unique global minimum.

Thus, the optimization problem for value-oriented aggregation is strictly convex with a unique solution for $p > 1$, and can be solved using standard convex optimization techniques. For $p = 1$ it's proven to be convex, but not strictly, which means that in the worst case scenario there could theoretically be several nearby optimal solutions which all correspond to the same minimum objective function.

To demonstrate that U_p^A is convex and has a unique solution, we consider similar points as before:

- **Objective Function Convexity:**

- The ℓ_p -norm $\|\cdot\|_p$ is a convex function for $p \geq 1$.
- The matrices $\mathbf{D}^+ = \Delta \mathbf{W}^T \mathbf{G}^+$ and $\mathbf{D}^- = \Delta \mathbf{W}^T \mathbf{G}^-$ are affine transformations of $\mathbf{w}^{s'}$.
- The absolute value function is convex.
- The composition of convex functions (absolute value and ℓ_p -norm) with affine transformations is convex.

- **Constraints Convexity:**

- The constraints $\sum_{i=1}^{|V|} w_i^{s'} = 1$ and $w_i^{s'} \geq 0$ define a simplex, which is a convex set.

- **Strict Convexity and Uniqueness:**

- For $p > 1$, the ℓ_p -norm is strictly convex.
- Therefore, the objective function is strictly convex, ensuring a unique global minimum.

Thus, the optimization problem for action-oriented aggregation is strictly convex with a unique solution for $p > 1$, and can be solved using standard convex optimization techniques. Similar analysis than before for $p = 1$. Both methods are now proven to yield a unique solution and we can confidently say that VOA and AOA satisfy the commutativity property 3.3 because the adding distances is a commutative operation and both methods will always yield the same solution regardless of the order of the agents. Also, the uniqueness of the solution ensures that AOA is the only method providing an aggregation function based on the Value Utility Alignment Principle 3.5.

4.3.3 Numerical Solution

To solve the minimization problem, we used the BFGS (Broyden–Fletcher–Goldfarb–Shanno) optimization algorithm via the `scipy.optimize.minimize` function in Python. BFGS is a quasi-Newton method that approximates the Hessian matrix to optimize efficiently without

needing explicit second-order derivatives. The algorithm iteratively updates the consensus weight vector \mathbf{w}^S by computing the gradient of the objective function U_p^V and determining the optimal step size via a line search. The Hessian approximation improves with each iteration, allowing for rapid convergence, especially in smooth optimization landscapes. Constraints, such as ensuring the weights sum to one ($\sum_{i=1}^{|V|} w_i^S = 1$) and non-negativity ($w_i^S \geq 0$), are handled directly by the `scipy.optimize.minimize` function.

4.4 Aggregation function analysis: are VOA and AOA the same?

The two different aggregation methods have been presented, and it has been shown that they converge to an optimal solution. Despite being grounded in the different principles of minimizing weight differences and minimizing action judgement differences, it is still to be explored whether they produce different results in practice. Both methods minimize their respective aggregation functions producing the optimal weight for each value, ensuring that they sum 1 (i.e. ensuring it conforms a value system). As such, the output weights of each methods can be compared in order to evaluate how different their results actually are. Later, I will also provide a reasoning of the differences of the methods providing a toy example explanation.

In order to start comparing both methods, I will consider two different aspects. The first is **the ordering of value preferences**, which refers to the ranking of values by their importance to the agent. This has traditionally been relevant in social choice and preference aggregation because general models of preferences often do not account for how much more one option is preferred over another. In this context, weighted preferences provide a more specific representation and can be used to infer a general ranking by comparing their numerical values. Comparing whether two orderings agree is generally straightforward—one simply orders the list from the highest to the lowest weight. However, some weights may be very close, and due to numerical errors in our optimization methods, the ordering might be interpreted as different when it is essentially the same. To address this, in our comparison of the models across different setups, we use a Python function that compares rankings while allowing for some tolerance. A strict True (1.0) or False (0.0) comparison of orderings is considered.

Besides comparing the differences between the ranking of value preferences of the two methods, I will also consider **significant differences in absolute weights** for any value. If, for any value, the weights differ by more than $\frac{1}{|V|} \cdot 20\%$, this value will be considered significantly different between the two methods. This means that the weights should not differ by more than one fifth of the evenly distributed weight value. For example, for five values ($|V| = 5$), if the weights were evenly distributed, we would have $w_i = 0.20 \forall i$, and a value i would be considered significantly different if $|w_i - w'_i| \geq 0.04$ (i.e., $0.20 \cdot 20\%$).

Let's consider an example. First, we need to choose the number of agents ($|H|$), values ($|V|$), and actions ($|A|$) in a given context. Then, we select the weights w_i^k for each agent and value, encapsulated in the matrix W . Next, we select the judgments from each value to each action (v^+, v^-), encapsulated in the matrices G^+ and G^- . Finally, we choose the desired value of p . In this first example, we select 4 agents, 5 actions, and 5 values, with all matrices randomly chosen while satisfying the constraints.¹

Table 1 shows that the output of both aggregation methods is somewhat different. Besides $p = 2$ for which the weights are exactly the same (and thus their ordering), we see that all the weights are different for the other values of p ; although just for w_2^S and w_3^S when $p = 1$ the optimal weights differ more than the $\frac{1}{|V|} \cdot 20\%$ threshold. The orderings also differ for $p = 1$ and for $p = \infty$ meaning that for this example, and these two values of p , we get different consensus

¹To ensure that the weights for each agent sum to one, random numbers were uniformly selected from an interval and then normalized.

Value of p	Weight-oriented	Action-oriented
1	[0.2787, 0.1986, 0.1484, 0.2837, 0.0905] $w_4 \succ w_1 \succ w_2 \succ w_3 \succ w_5$	[0.2895, 0.0553, 0.2648, 0.3173, 0.0730] $w_4 \succ w_1 \succ w_3 \succ w_5 \succ w_2$
2	[0.3075, 0.1475, 0.1950, 0.2325, 0.1175] $w_1 \succ w_4 \succ w_3 \succ w_2 \succ w_5$	[0.3075, 0.1475, 0.1950, 0.2325, 0.1175] $w_1 \succ w_4 \succ w_3 \succ w_2 \succ w_5$
∞	[0.3181, 0.1581, 0.2056, 0.1900, 0.1281] $w_1 \succ w_3 \succ w_4 \succ w_2 \succ w_5$	[0.3034, 0.1513, 0.2054, 0.2250, 0.1149] $w_1 \succ w_4 \succ w_3 \succ w_2 \succ w_5$

Table 1: Comparison of optimal weights and their ordering for different aggregation methods with random matrices. The elements in red show differences in the results of both methods for a given value of p . Weights are in red when they differ more than 0.04. Orderings are in red if they are not the same. The seed used to generate pseudo-random numbers is ‘numpy.random.seed(20)’

preferences by using these two aggregation methods.

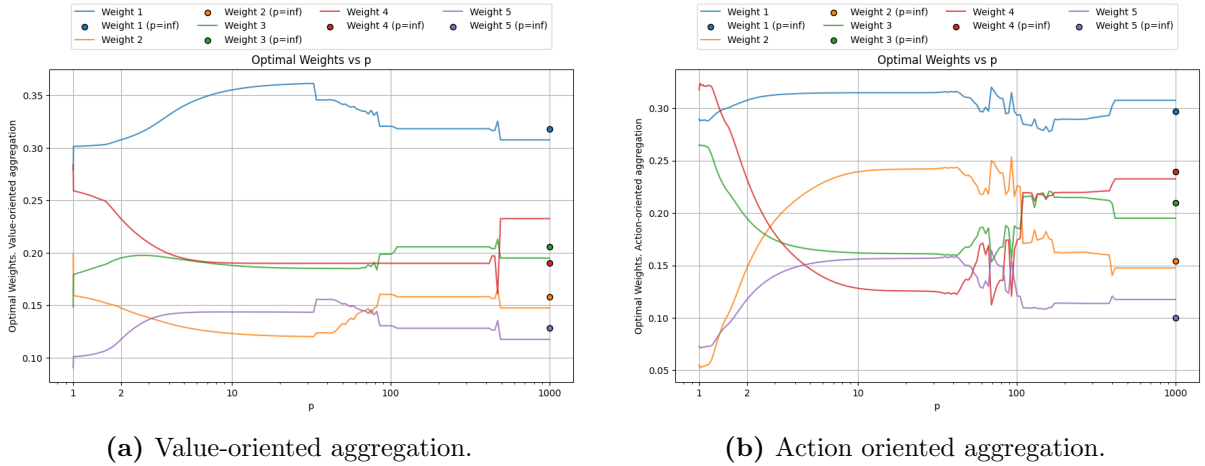


Figure 2: Consensus weights for both aggregation methods from $p = 1$ to $p = \infty$.

You can visualize the optimal weights for different values of p for both aggregation methods in Figure 2. The evolution of some weights like w_1 and w_5 is similar, while others like w_2 and w_3 show different behaviours in both graphs. Note that for both aggregation methods yield the same consensus weights for $p = 2$, but for other values like $p = 10$ there are big differences in terms of absolute weights and orderings. Also, there is a sudden change for all weights between $p = 400$ and $p = 500$ which reflects the limit of numerical underflow i.e. where very high exponents for positive numbers smaller than one mean that just zero is stored. Thus, both images should just consider values of $p < 400$ and then provide the $p = \infty$ to give a better perspective of the real tendency of the methods.²

In order to explore these differences further besides this specific example with 4 agents, 5 values and 5 actions, a new experiment was conducted. The optimal weights between both methods were calculated and compared for different values of $|V|, |A|, |H|, p$. Ten different values of $|A|$ are explored (from $|A| = 1$ to $|A| = 10$), and the results are averaged out in the six grids below for every combination of number of values $|V|$ and number of agents $|H|$. As such, Table 2

²301 solutions with different values of p were considered to create these plots. The values of p were chosen using a logarithmic scale, with 150 values between $p = 1$ and $p = 2$, and another 150 between $p = 2$ and $p = 1,000$, plus one value for $p = \infty$. All 301 solutions were calculated for both methods in less than 2 minutes.

can be read as "What percentage of setups with this $|V|$ and this $|H|$ resulted in the same ordering for both methods?". Table 3, instead, focuses on the significant differences between the consensus weights produced by both methods; it also averages over $|A|$.

$p = 1$						$p = 2$						$p = \infty$					
$ H \backslash V $	2	3	4	5	10	$ H \backslash V $	2	3	4	5	10	$ H \backslash V $	2	3	4	5	10
2	1.0	1.0	1.0	1.0	1.0	2	1.0	1.0	1.0	1.0	1.0	2	1.0	1.0	1.0	1.0	1.0
3	1.0	0.8	0.6	0.7	0.7	3	1.0	1.0	1.0	1.0	1.0	3	1.0	0.9	0.8	0.7	0.7
4	1.0	0.4	0.3	0.3	0.8	4	1.0	1.0	1.0	1.0	1.0	4	1.0	0.7	0.7	0.4	0.5
5	1.0	0.0	0.4	0.2	0.0	5	1.0	1.0	1.0	1.0	1.0	5	1.0	0.5	0.6	0.3	0.0
10	1.0	0.0	0.0	0.0	0.0	10	1.0	1.0	1.0	1.0	1.0	10	1.0	0.0	0.0	0.0	0.3

Table 2: Percentage of setups with the same ordering for both methods.

$p = 1$						$p = 2$						$p = \infty$					
$ H \backslash V $	2	3	4	5	10	$ H \backslash V $	2	3	4	5	10	$ H \backslash V $	2	3	4	5	10
2	0.0	0.0	0.0	0.0	0.0	2	0.0	0.0	0.0	0.0	0.0	2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.1	0.4	0.2	0.0	3	0.0	0.0	0.0	0.0	0.0	3	0.0	0.0	0.0	0.0	0.2
4	0.0	0.7	0.7	1.1	0.2	4	0.0	0.0	0.0	0.0	0.0	4	0.0	0.1	0.4	0.4	0.7
5	0.0	1.9	0.8	1.6	1.4	5	0.0	0.0	0.0	0.0	0.0	5	0.0	0.4	0.3	0.7	1.1
10	0.0	5.1	2.8	5.3	3.8	10	0.0	0.0	0.0	0.0	0.0	10	0.0	0.9	0.9	1.8	2.7

Table 3: Average of significant differences for both methods.

Both tables suggest that for $p = 2$, $|V| = 2$ and $|H| = 2$ both value aggregation methods work pretty similar. In the next section, we will prove mathematically that this is case. Also, we note more differences in ordering and absolute weights appear as $|V|$ and $|H|$ get bigger, suggesting that the methods yield clearly different consensus weights (if $p \neq 2$). It is also true that adding more values to a decision can results in more ways of the methods being different because it's more likely that the ordering is the same for 3 values than for 10.

4.5 Analytical solution: conditions and special cases.

In this subsection, we will prove mathematically that both aggregation methods yield equivalent solutions for the following three edge cases. Additionally, this subsection works as an attempt to explore the proposed aggregation functions analytically, instead of numerically.

4.5.1 First Order Necessary Conditions

For the three proofs, we will use the First Order Necessary Conditions (FONCs) of both objective functions, U_p^V and U_p^A . The FONCs refer to the equations that the consensus weights \mathbf{w}^S must satisfy in order for the derivative of the objective function to be zero. Since both objective functions are convex, the FONCs are also sufficient conditions for finding the minimum, as there is only one set of weights that can satisfy this condition. Note that using FONCs is the approach to solving the unconstrained optimization problem. However, in all three cases explored in this subsection, the set of weights that satisfy the FONCs also satisfy the constraints, simplifying the proofs.

Recall that U_p^V and U_p^A are multivariate distance functions whose values depend solely on the

candidate weights $\mathbf{w}^{s'}$ of the value system we are optimizing. All the input weights w_i^k for every agent k and every value i are provided in the matrix W . The goal is to find the set of candidate weights $\mathbf{w}^{s'}$ that minimize the objective function (i.e., satisfy the FONCs); these minimizing weights are called the consensus weights, denoted as $\mathbf{w}^S = [w_i^S, w_2^S, \dots, w_{|V|}^S]$.

For the U_p^V objective distance function,

$$U_p^V = \left(\sum_{k=1}^{|H|} \sum_{i=1}^{|V|} |w_i^k - w_i^{s'}|^p \right)^{1/p} = u^{1/p}$$

To find the minimum, we apply the FONC by differentiating with respect to each candidate weight $w_i^{s'}$ and setting the derivative to zero. For a solution to be the minimum, all the following equations must hold for $i' \leq |V|$:

$$\frac{\partial U_p^V}{\partial w_{i'}^S} = \frac{1}{p} \cdot u^{\frac{1-p}{p}} \left[\sum_{k=1}^{|H|} -p \cdot |w_{i'}^k - w_{i'}^{s'}|^{p-1} \cdot \text{sign}(w_{i'}^k - w_{i'}^{s'}) \right] = 0$$

The scalars $-p$ and $1/p$ can be ignored for the solution by dividing them out. Additionally, since u is always positive (as it represents a distance function) and raised to a negative power (i.e., it is in the denominator), it can also be ignored for the solution. The $\text{sign}()$ function appears due to the derivative of the absolute value. Therefore, after applying the FONCs for a general p , we know that the minimum of the U_p^V objective distance function must satisfy the following simultaneous equations for each value:

$$\sum_{k=1}^{|H|} |w_{i'}^S - w_{i'}^k|^{p-1} \cdot \text{sign}(w_{i'}^k - w_{i'}^S) = 0 \quad (1)$$

Now, for the U_p^A objective distance function, I will assume for all the proofs that the negative judgement function $J_{s'}^-(a_j)$ is zero to make the proofs more clear, because the methods are equivalent and shorter. Also, remember for this whole subsection the formula of the global judgement function (or utility) of an agent $J^+(a_j) = \sum_{i=1}^{|V|} (w_i \cdot v_i^+(a_j))$.

$$U_p^A = \left[\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} |J_k^+(a_j) - J_S^+(a_j)|^p \right]^{1/p} = u^{1/p}$$

To find the minimum, again, we apply the FONC by differentiating with respect to each candidate weight $w_i^{s'}$ and setting the derivative to zero. For a solution to be the minimum, all the following equations must hold for $i' \leq |V|$:

$$\frac{\partial U_p^A}{\partial w_{i'}^S} = \frac{1}{p} \cdot u^{\frac{1-p}{p}} \cdot \left[\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} -p \cdot |J_k^+(a_j) - J_S^+(a_j)|^{p-1} \cdot \text{sign}(J_k^+(a_j) - J_S^+(a_j)) \cdot v_{i'}^+(a_j) \right] = 0$$

As before, the scalars and the u term can be ignored for the solution. The $\text{sign}()$ function appears due to the derivative of the absolute value. The $v_{i'}^+(a_j)$ term appears when taking the derivative of $J_S^+(a_j)$ with respect to $w_{i'}^{s'}$, as it is the only term multiplying $w_{i'}^{s'}$. Therefore, after applying the FONC for a general p , we know that the minimum of the U_p^V objective distance function must satisfy the following simultaneous equations for each value:

$$\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} |J_k^+(a_j) - J_S^+(a_j)|^{p-1} \cdot \text{sign}(J_k^+(a_j) - J_S^+(a_j)) \cdot v_{i'}^+(a_j) = 0. \quad (2)$$

4.5.2 Mathematical Derivation of Consensus Weights for $p = 2$

Here, we show that both VOA and AOA methods yield the same consensus weights for $p = 2$. First, we proof that U_2^V consensus weights correspond to the average weights of all agents. Then, we proof that average weights are also the consensus weights for AOA. Finally, we check that the average weights satisfy the constraints and are thus a reasonable solution for the aggregation problem.

Theorem 4.1. *The consensus weights for U_2^V are the average weights for all agents.*

Proof. We begin by applying the First Order Necessary Condition (FONC) from Equation 1, and note that for $p = 2$, the exponent $p - 1$ becomes 1, and the product of the absolute value times the sign() function simplifies to the identity. The FONC equation for $p = 2$ is:

$$\sum_{k=1}^{|H|} \left| w_{i'}^S - w_{i'}^k \right|^{p-1} \cdot \text{sign} \left(w_{i'}^k - w_{i'}^S \right) = 0$$

Which simplifies to:

$$\sum_{k=1}^{|H|} (w_{i'}^k - w_{i'}^S) = 0.$$

We can extract the consensus weight as it is not affected by the summation:

$$\sum_{k=1}^{|H|} w_{i'}^k = |H| \cdot w_{i'}^S,$$

which leads to the solution for the consensus weight $w_{i'}^S$:

$$w_{i'}^S = \frac{1}{|H|} \sum_{k=1}^{|H|} w_{i'}^k.$$

Thus, the consensus weights for U_2^V are the average of the weights for all agents. □

Theorem 4.2. *The consensus weights for U_2^A are the average weights for all agents.*

Proof. We begin by applying the First Order Necessary Condition (FONC) from Equation 2, and note that for $p = 2$, the exponent $p - 1$ becomes 1, and the product of the absolute value times the sign() function simplifies to the identity. The FONC equation for $p = 2$ is:

$$\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} \left| J_k^+(a_j) - J_S^+(a_j) \right|^{p-1} \cdot \text{sign} \left(J_k^+(a_j) - J_S^+(a_j) \right) \cdot v_{i'}^+(a_j) = 0.$$

Which simplifies to:

$$\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} (J_k^+(a_j) - J_S^+(a_j)) \cdot v_{i'}^+(a_j) = 0.$$

Expanding $J^+(a_j) = \sum_{i=1}^{|V|} (w_i \cdot v_i^+(a_j))$ and taking common factor:

$$\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} v_{i'}^+(a_j) \cdot \sum_{i=1}^{|V|} \left(v_i^+(a_j) \cdot (w_i^k - w_i^S) \right) = 0$$

Now, by sliding the $\sum_{k=1}^{|H|}$ summation term to the right we get the final equation:

$$\frac{\partial U_2^A}{\partial w_{i'}^S} = 0 \implies \sum_{j=1}^{|A|} v_{i'}^+(a_j) \cdot \sum_{i=1}^{|V|} v_i^+(a_j) \cdot \sum_{k=1}^{|H|} (w_i^k - w_i^S) = 0$$

As this is a convex minimization problem, there is just one set of weights that satisfy the FONCs (which is equivalent to the equation above in this case) and correspond to the minimum of U_2^A . If we try the average weights solution, the final summation $\sum_{k=1}^{|H|} (w_i^k - w_i^S)$ becomes zero and thus satisfies the FONC, meaning that the average weights are also a solution for U_2^A . \square

Lemma 4.3 (Bounds). *If all given weights w_i^k for each value i and agent k are within the interval $[0, 1]$, then the average consensus weight $w_i^S = \frac{1}{|H|} \sum_{k=1}^{|H|} w_i^k$ is also within $[0, 1]$.*

Proof. Since each $w_i^k \in [0, 1]$, we have:

$$0 \leq w_i^k \leq 1 \quad \text{for all } k = 1, 2, \dots, |H|.$$

Taking the average:

$$0 = \frac{1}{|H|} \sum_{k=1}^{|H|} 0 \leq \frac{1}{|H|} \sum_{k=1}^{|H|} w_{i'}^k \leq \frac{1}{|H|} \sum_{k=1}^{|H|} 1 = 1.$$

Thus, the consensus weight $w_i^S \in [0, 1]$, as required. \square

Lemma 4.4 (Constraint). *If, for each agent k , the weights $w_1^k, w_2^k, \dots, w_{|V|}^k$ sum to one, i.e., $\sum_{i=1}^{|V|} w_i^k = 1$, then the average of these weights across all agents for each value also sums to one: $\sum_{i=1}^{|V|} w_i^S = 1$.*

Proof. Consider the consensus weight for each value i , defined as the average:

$$w_i^S = \frac{1}{|H|} \sum_{k=1}^{|H|} w_i^k.$$

Summing over all values:

$$\sum_{i=1}^{|V|} w_i^S = \sum_{i=1}^{|V|} \frac{1}{|H|} \sum_{k=1}^{|H|} w_i^k = \frac{1}{|H|} \sum_{k=1}^{|H|} \sum_{i=1}^{|V|} w_i^k = \frac{1}{|H|} \sum_{k=1}^{|H|} 1 = \frac{|H|}{|H|} = 1.$$

Thus, the consensus weights w_i^S also satisfy the constraint that their sum equals one. \square

Thus, we have shown that the average weight is the solution for the unconstrained and unbounded minimization problem, and that it satisfies the constraints and bounds. Hence, it is the solution of VOA and AOA for $p = 2$.

Note in Section 4.2.1 we presented the that median, the mean and the mid-range are the aggregation metrics that minimize the p -norm of the residuals for $p = 1$, $p = 2$ and $p = \infty$ respectively, for the one-dimensional case. Now we have seen that for $p = 2$ this is still the case. However, it is not necessary for the meadian and the mid-range of the weights to satisfy the constraints³. So while there solutions will tend to the median and midrange (for the preferences over values in VOA and the utilities over actions in AOA) for $p = 1$ and $p = \infty$ as we will see on Section 4.6, this is not necessarily the solution in all cases due to the constraints.

³Imagine the scenario like in Section 4.6 but without constraints. We got 3 agents h_k and 3 values v_i where each agent gives a total weight to a different value so that $w_i^k = 1$ if $k = i$ and $w_i^k = 0$ otherwise. Then the median set of weights is $w^{s'} = [0.0, 0.0, 0.0]$ and the mid-range set of weights is $w^{s'} = [0.5, 0.5, 0.5]$, none of which satisfy the constraints.

4.5.3 Mathematical Derivation of Consensus Weights for $|V| = 2$

Theorem 4.5. *For $|V| = 2$, both the value-oriented and action-oriented aggregation methods yield identical consensus weights.*

Proof. Using the first-order conditions for U_p^V on Equation 1 we get that the optimal weights w_1^S and w_2^S for the two values in this case should satisfy the following simultaneous equations:

$$\sum_{k=1}^{|H|} \left| w_1^k - w_1^S \right|^{p-1} \cdot \text{sign} \left(w_1^k - w_1^S \right) = 0. \quad (3)$$

$$\sum_{k=1}^{|H|} \left| w_2^k - w_2^S \right|^{p-1} \cdot \text{sign} \left(w_2^k - w_2^S \right) = 0. \quad (4)$$

Note that for a solution w_1^S to Equation 3, we get that $w_2^S = 1 - w_1^S$ is also a solution to Equation 4. As there is only one solution, we know that the solution to the unconstrained minimization problem of U_p^V does indeed satisfy the constraint. Now, consider the first-order condition for U_p^A from Equation 2:

$$\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} \left| J_k^+(a_j) - J_S^+(a_j) \right|^{p-1} \cdot \text{sign} \left(J_k^+(a_j) - J_S^+(a_j) \right) \cdot v_{i'}^+(a_j) = 0.$$

Focusing on the term $|J_k^+(a_j) - J_S^+(a_j)|$ and noting we got just two values here, we expand it as

$$\left| J_k^+(a_j) - J_S^+(a_j) \right| = \left| v_1^+(a_j)(w_1^k - w_1^S) + v_2^+(a_j)(w_2^k - w_2^S) \right|.$$

We know that $w_1^k + w_2^k = 1$ for all k because there are just two values. Now, we assume that the consensus weights also satisfy the constraints so that $w_1^S + w_2^S = 1$. Now, we can substitute any w_2 from the expression by $1 - w_1$, which simplifies to:

$$\left| J_k^+(a_j) - J_S^+(a_j) \right| = \left| (v_1^+(a_j) - v_2^+(a_j)) \cdot (w_1^k - w_1^S) \right|.$$

Substituting this expression back into Equation 2 (which is also shown above), we obtain:

$$\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} \left| (v_1^+(a_j) - v_2^+(a_j)) \cdot (w_1^k - w_1^S) \right|^{p-1} \cdot \text{sign} \left((v_1^+(a_j) - v_2^+(a_j)) \cdot (w_1^k - w_1^S) \right) \cdot v_{i'}^+(a_j) = 0.$$

Here, we can split the $(v_1^+(a_j) - v_2^+(a_j)) \cdot (w_1^k - w_1^S)$ product and then slide the summation signs as follows:

$$\sum_{j=1}^{|A|} \left| v_1^+(a_j) - v_2^+(a_j) \right|^{p-1} \cdot \text{sign} \left(v_1^+(a_j) - v_2^+(a_j) \right) \cdot v_{i'}^+(a_j) \cdot \sum_{k=1}^{|H|} \left| w_1^k - w_1^S \right|^{p-1} \cdot \text{sign} \left(w_1^k - w_1^S \right) = 0.$$

Notice that the above condition is satisfied if

$$\sum_{k=1}^{|H|} \left| w_1^k - w_1^S \right|^{p-1} \cdot \text{sign} \left(w_1^k - w_1^S \right) = 0,$$

which aligns with the first-order condition of the value-oriented aggregation method. Since both aggregation methods are convex and possess a unique solution, we conclude that for $|V| = 2$, the consensus weights obtained via both methods are identical. \square

4.5.4 Mathematical Derivation of Consensus Weights for $|H| = 2$

Lemma 4.6. *For $|H| = 2$, both the value-oriented and action-oriented aggregation methods yield identical consensus weights.*

Proof. We first establish that for two agents, the optimal weights for the value-oriented aggregation method are the averages of the weights of the two agents. Observe the first-order condition for U_p^V on Equation 1:

$$\sum_{k=1}^{|H|} |w_{i'}^S - w_{i'}^k|^{p-1} \cdot \text{sign}(w_{i'}^k - w_{i'}^S) = 0.$$

Since $|H| = 2$, we can separate the sum as follows:

$$|w_{i'}^S - w_{i'}^1|^{p-1} \cdot \text{sign}(w_{i'}^1 - w_{i'}^S) = -|w_{i'}^S - w_{i'}^2|^{p-1} \cdot \text{sign}(w_{i'}^2 - w_{i'}^S).$$

Given that the absolute value terms are positive, the sign expressions must be opposite:

$$\text{sign}(w_{i'}^1 - w_{i'}^S) = -\text{sign}(w_{i'}^2 - w_{i'}^S).$$

This implies that one of the weights $w_{i'}^k$ is greater than $w_{i'}^S$ and the other is smaller. Consequently, we have:

$$|w_{i'}^S - w_{i'}^1|^{p-1} = |w_{i'}^S - w_{i'}^2|^{p-1},$$

which simplifies to

$$|w_{i'}^S - w_{i'}^1| = |w_{i'}^S - w_{i'}^2|.$$

Considering the opposite signs, we obtain:

$$w_{i'}^S - w_{i'}^1 = w_{i'}^2 - w_{i'}^S,$$

which leads to

$$w_{i'}^S = \frac{1}{2} (w_{i'}^1 + w_{i'}^2).$$

Thus, the optimal weights for the value-oriented aggregation method are the averages of the weights of the two agents. Next, consider the first-order condition for U_p^A from Equation 2:

$$\sum_{k=1}^{|H|} \sum_{j=1}^{|A|} |J_k^+(a_j) - J_S^+(a_j)|^{p-1} \cdot \text{sign}(J_k^+(a_j) - J_S^+(a_j)) \cdot v_{i'}^+(a_j) = 0.$$

Define $D_k^+(a_j) = J_k^+(a_j) - J_S^+(a_j) = \sum_{i=1}^{|V|} v_i^+(a_j) \cdot (w_i^k - w_i^S)$. For $|H| = 2$, we can split the sum as:

$$\sum_{j=1}^{|A|} |D_1^+(a_j)|^{p-1} \cdot \text{sign}(D_1^+(a_j)) \cdot v_{i'}^+(a_j) = - \sum_{j=1}^{|A|} |D_2^+(a_j)|^{p-1} \cdot \text{sign}(D_2^+(a_j)) \cdot v_{i'}^+(a_j).$$

To satisfy this first-order condition, it suffices to show that for every action a_j the following holds:

$$|D_1^+(a_j)|^{p-1} \cdot \text{sign}(D_1^+(a_j)) = -|D_2^+(a_j)|^{p-1} \cdot \text{sign}(D_2^+(a_j)).$$

Following similar reasoning as before, the signs must be opposite:

$$\text{sign}(D_1^+(a_j)) = -\text{sign}(D_2^+(a_j)),$$

and thus

$$|D_1^+(a_j)|^{p-1} = |D_2^+(a_j)|^{p-1},$$

which implies

$$D_1^+(a_j) = -D_2^+(a_j).$$

This condition holds for every action a_j if the weights for a value are the averages of the weights of the two agents for that value. This can be verified by noting that:

$$w_i^1 - w_i^S = -(w_i^2 - w_i^S),$$

and expanding $D_k^+(a_j)$:

$$\sum_{i=1}^{|V|} v_i^+(a_j) \cdot (w_i^1 - w_i^S) = - \sum_{i=1}^{|V|} v_i^+(a_j) \cdot (w_i^2 - w_i^S).$$

Therefore, both aggregation methods yield identical consensus weights for $|H| = 2$. \square

4.6 Toy Example: Urbanism

The aim of this section is to show the theoretical differences between value-oriented and action-oriented aggregation methods through the analysis of a toy example. In most cases, there is a simple final explanation for why do the methods differ for a specific value of p . However, the fact that they differ in the final value of consensus weights or even the ordering of the values, showcases a fundamental decision to make when aggregating value systems: prioritize similar value value preferences or similar judgement of actions. In the final Section 5, possible approaches to make this decision are proposed.

In order to explore the differences between VOA and AOA, we need a scenario with at least 3 agents and at least 3 values, if not, both methods would yield the same consensus weights as we have shown in Section 4.5. Thus, we choose the simplest case with 3 agents, 3 values and 1 action. To make the analysis simple, we set the value weights of each agent so that they correspond to fully endorse a value so that is $w_i^k = 1$ if $k = i$, and $w_i^k = 0$ if $k \neq i$; i.e. W is the 3×3 identity matrix. Before specifying the judgement functions (v^+, v^-) which indicate how values relate to actions, we can already perform the value-oriented aggregation which results in equal weights $w^S = [0.33, 0.33, 0.33]$ for every p .

Now, let's consider the context of urbanism and the values of individual freedom (w_1), environmentalism (w_2) and transport efficiency (w_3). The only action a_1 to be considered is "commuting in private vehicle". The three values judge the performance of action a_1 as follows (simple numbers are chosen in this example). Individual freedom judges a_1 it's a positive thing $v_1^+(a_1) = 1$. For environmentalism it's a negative thing $v_2^+(a_1) = 1$. And imagine the available infrastructure is quite poor and cannot accommodate all commuters, so that in terms of transport efficiency commuting by private vehicle is a damaging action: $v_3^+(a_1) = -1$. When it comes to G^- we can argue that these judgement is 0 for all values because not doing the action (commuting by private vehicle, in this case) does not mean that the alternative action chosen will necessarily promote or demote any of the values.

Note that the G matrix already represents the global judgements of the agents to the action a_1 because just the judgement of one weight is important to each of them. In order to understand the evolution of the weights for the action-oriented aggregation over the different values of p , one should bear in mind the motivation of the method. By minimizing U_p^A , the distance between the consensus and the agents' judgement of actions is minimized according to the p -norm. As such, the consensus weights are chosen to minimize the judgement distances.

We now employ the analysis of the different edge cases of the p -norm presented in Section 2.3 to understand the action-oriented aggregation:

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad G^+ = \begin{bmatrix} +1 \\ -1 \\ -1 \end{bmatrix}$$

	$p = 1$	$p = 2$	$p = \infty$
w_1^S	0.00	0.33	0.50
w_2^S	0.50	0.33	0.25
w_3^S	0.50	0.33	0.25

Figure 3: Weight matrix, positive judgement vector, and consensus weights of the action-oriented aggregation methods for different p values.

- For $p = 1$, the judgement of the consensus value system should be the median of the judgements because that minimizes the absolute distances, which is -1 in this case. The weights are chosen accordingly, with $w_1^S = 0$ so that the final global judgement J_S^+ doesn't include the individual freedom judgement v_1^+ at all. This way, the global judgement of the consensus value system corresponds with the median judgement $G_S^+(a_1) = -1$. These consensus weights maximize the social utility of the judgement of actions.
- For $p = 2$, the judgement of the consensus value system should be the average of the judgements because that minimizes the squared distances. The weights are chosen accordingly, with each consensus weight being the average weight of the agents for each value v_i the consensus weight is: $w_i^S = \frac{1}{3} \sum_{k=1}^3 w_i^k$. The mean judgement is $G_S^+(a_1) = -0.33$ and as shown in Section 4.5.2, because for $p = 2$, average consensus weights imply average consensus judgement. These consensus weights correspond to the egalitarian approach of the judgement of actions.
- For $p = \infty$, the judgement consensus value system should be the mid-range —the average between the maximum and the minimum— of the judgements because that minimizes the maximum absolute distance. The weights are chosen accordingly so that the consensus judgement function is $G_S^+(a_1) = 0$, which is the mid-range of the judgements. As such, individual freedom gets $w_2^S = 0.50$ in order to compensate for environmentalism and transport efficiency $w_1^S = w_3^S = 0.25$ which judge the action negatively. These consensus weights maximize the fairness of the judgement of actions.

This analysis highlights the different principles guiding both aggregation methods. On the one hand, value-oriented aggregation is based on the principle of getting a consensus value system with similar preferences of values, which correspond to the weights in this case. On the other, the action-oriented aggregation is based on the principle of getting a consensus value system with similar judgement of actions, which is characterized by the the global judgement functions (J^+, J^-) . The different principles of the two methods lead to different outputs of aggregation. While VOA considers producing the equal weights solution as the most utilitarian ($p = 1$) and fair ($p = \infty$) solution, AOA considers that utility and fairness are properties of the final judgement of actions, not of the value preferences which leads to different results as shown in Figure 3

Usually, both approaches will produce different results, we will choose one or the other depending on how we plan to use the aggregated value system. VOA is a general aggregation method suitable for most applications, however in applications where the consequences of the value system are important (in terms of making decisions or selecting an action) AOA may be more appropriate. This is explored further in Section 5.1.

4.7 European Value Study data

In this section, I aim to infer data about value systems from a large-scale survey in order to further exemplify the differences between both methods. Due to the lack of value-related data, we resort to one of the only surveys studying people’s values. However the reader should be aware that this study does not follow the precise framing of values described in the methodology section, so this section should be understood as a proof of concept.

The European Values Study (EVS) (Study 2020) is a cross-national survey research program on basic human values. It has been repeated every 9 years since 1981 and it provides insights into the ideas, beliefs, preferences, attitudes, values and opinions of citizens all over Europe. In (Lera-Leri et al. 2024), this data is used to infer the preferences and understandings for two values: *religiosity* (w_1) and *permissiveness* (w_2), and for the actions of *adoption by homosexual couple* (a_1) and *divorce* (a_2). In order to compare the two aggregation methods we need more than two values, so we also consider *liberalism* (w_3) and *conservatism* (w_4). Due to the limited data in the survey, we consider the G^- matrix to be 0, although we could have assumed it is the opposite of the G^+ matrix which would lead to a similar result.

For our example we have taken the 2017 wave (Gedeshi et al. 2017), which is the most recent. This dataset includes data from around 60.000 respondents from 36 European countries, which become 50.000 after removing the ones with missing values for the questions we require. Each country has its own weights for each of the four values, and judgements for the actions a_1 and a_2 from the previous paragraph. Our aim is to aggregate each of the 36 country value systems into a common one. The weights for permissiveness (w_1) and religiosity (w_2) are extracted from the their proportion of people in a country that considered religion important in their life (Question Q1F). The weights for liberalism (w_3) and conservatism (w_4) are extracted for the proportion of people who chose left (1,2,3 or 4) over right (7,8,9,10) in a 1 to 10 left-right spectrum (Q31). These proportions for the four weights of each agent are halved so that the weights add up to one. The judgement functions of each value towards each of the actions are taken by averaging the opinions of each group (permissive, religious, conservative, and liberal) with their opinions to divorce (Q44G: "Can divorce be always justified, never justified, or something in between?") and adoption by homosexual couples (Q27A: "How much do you agree or disagree with the statement: Homosexual couples are as good parents as other couples?").

	Religiousness (w_1)	Permissiveness (w_2)	Liberalism (w_3)	Conservatism (w_4)
Divorce (a_1)	-0.01	0.51	0.37	0.16
H. Parenting (a_2)	-0.29	0.20	0.11	-0.17

Table 4: V matrix. Judgement of actions by the different values based on European average responses on each group.

With the W and G^+ matrices, we can now perform both aggregation methods (see Table 6). For $p = 1$, we see that both methods yield different consensus weights for each value although they not differ significantly. We see an ordering change between the values of liberalism and permissiveness, but they have similar weights in both consensus so it does not represent an actual change in preferences. For $p = 2$ we get exactly the same weights as shown in 4.5.2. For $p = \infty$ we observe very significant changes in the consensus weights of liberalism and permissiveness which in turn alters most part of the ordering.












Country	Religiousness (w_1)	Permissiveness (w_2)	Liberalism (w_3)	Conservatism (w_4)
 CZ	0.10	0.40	0.23	0.28
 DK	0.12	0.39	0.27	0.23
 FR	0.17	0.33	0.28	0.22
 GE	0.47	0.03	0.12	0.38
 DE	0.15	0.34	0.35	0.15
 IT	0.32	0.18	0.22	0.28
 ME	0.42	0.08	0.29	0.21
 PL	0.41	0.09	0.12	0.38
 RS	0.39	0.11	0.26	0.24
 ES	0.19	0.31	0.33	0.17
 GB	0.18	0.32	0.27	0.23

Table 5: W matrix. Weights of Religiousness, Permissiveness, Liberalism, and Conservatism for 11 out of the 36 different European countries used for the aggregation.

Value of p	Value-oriented	Action-oriented
1	[0.2600, 0.2399, 0.2400, 0.2598] $w_1 \succ w_4 \succ w_3 \succ w_2$	[0.2797, 0.2297, 0.2273, 0.2630] $w_1 \succ w_4 \succ w_2 \succ w_3$
2	[0.2603, 0.2394, 0.2369, 0.2631] $w_4 \succ w_1 \succ w_2 \succ w_3$	[0.2603, 0.2394, 0.2369, 0.2631] $w_4 \succ w_1 \succ w_2 \succ w_3$
∞	[0.2850, 0.2150, 0.2372, 0.2625] $w_1 \succ w_4 \succ w_3 \succ w_2$	[0.2569, 0.0898, 0.3409, 0.3121] $w_3 \succ w_4 \succ w_1 \succ w_2$

Table 6: Consensus weights of VOA and AAO for different values of p

5 Discussion and Conclusion

This project introduces two aggregation methods that address distinct aspects of the value aggregation problem, essential for AI alignment. The first, Value-Oriented Aggregation (VOA), focuses on aligning individual value preferences, while the second, Action-Oriented Aggregation (AOA), bridges the gap between value systems and decision-making by directly minimizing the divergence in action decisions. Both methods employ optimization techniques, notably p -norm minimization, to determine consensus outcomes. Overall, this research highlights two fundamentally different approaches to the value aggregation problem: one at the level of values (preferences) and the other at the level of actions (utility/judgement). These frameworks were validated using examples, including the European Value Study (EVS) dataset and a toy urbanism scenario, demonstrating their applicability to real-world contexts.

In contrast to the traditional ordinal ranking or pairwise comparison methods often found in the value alignment literature, this research introduces a novel framework based on weighted preferences. Although no ideal framework exists for human moral value preferences (each has its pros and cons), weighted preferences offer a clear advantage in the form of greater interpretability of the inputs and consensus value systems, allowing for a more nuanced understanding of how individual values influence decision-making. Furthermore, VOA focuses directly on weight differences and AOA employs the differences in the utility function which is a linear combination of judgment functions using also weight. The similarity of both methods was exploited to prove their equivalence in particular cases, such as when $p = 2$, or when there are two values $|V| = 2$

or two agents $|H| = 2$.

The aggregation methods developed here provide a versatile and general toolkit for decision-making processes, extending beyond the specific context of AI alignment. While the focus of this work was on value systems, its principles can be readily adapted to any decision-making scenario where multiple criteria or values must be considered like water resource managing or cultural heritage preservation. Additionally, these methods can be applied to the Lera-Leri et al. two-step optimization setup (Lera-Leri et al. 2022), even without each agent having the same understanding of values. As long as a global judgement function of actions is chosen out of the value preferences, the difference between the judgements can be calculated, a method following the Value Utility Alignment Principle (Section 3.5) can be defined. As such, this project also serves as a guideline to construct action-oriented value aggregation in different contexts.

5.1 Practical use cases: which value of p and which aggregation method?

Choosing the parameter p in our aggregation methods VOA and AOA requires careful consideration of practical implications. Setting $p = 1$ minimizes the overall dissatisfaction of individuals with respect to the consensus in terms of weights (VOA) or judgments (AOA), treating all individuals equally regardless of their proximity to the consensus. This is pretty useful for large-scale value aggregation where thousands or millions of agents’ value systems are considered and we aim to find the general tendency considering everyone equally. However, $p = 1$ may be overly restrictive since it invariably yields the median; consequently, the consensus value system remains largely unaffected by new input value systems unless they alter the median, and outliers exert no influence on the solution. For $p > 1$, the dissatisfaction of outliers with the consensus is accorded greater significance, which shifts the focus from the equal consideration for all agents and the utilitarian aim of reducing overall dissatisfaction. In many practical scenarios where diversity is crucial so that no agent is significantly misaligned with the consensus—such as policy-making in pluralistic societies, where the values of minority groups need to be adequately represented—a higher value of p , like $p = 2$ or greater, may be appropriate. The limiting case of $p = \infty$, despite adhering to the Rawlsian principle of justice by focusing exclusively on the most extreme positions, which could be useful depending on the problem at hand and on the decision maker’s priorities.

VOA and AOA serve distinct roles depending on the specific use case. Generally, if the primary goal to obtain an action prioritization and not so much value prioritization, AOA is preferred over VOA. VOA is especially suited when there is incomplete information about potential actions, or when these actions are not defined beforehand, such as in the aggregation of value preferences from donors of a non-profit or in AI alignment for open-world scenarios. In such contexts, VOA provides a flexible structure that accommodates a broad range of preferences without requiring concrete decision-making pathways. On the other hand, AOA becomes essential in more closed systems where the set of potential actions is well-defined and the consequences of those actions are the central concern, such as in participatory budgeting or disaster response decision-making. In these cases, AOA offers a more precise method for aligning value systems with final decisions, as it minimizes discrepancies between individual and collective utility on actions. While VOA provides a flexible solution for dynamic systems where actions are not clear yet, AOA delivers more accurate and actionable insights in well-defined decision-making environments.

5.2 Future Work

Value system inference, and particularly in value system aggregation are recent areas of research that still have room for improving current methods and many open problems. This project is no

exception, in particular, I would like to highlight the following points which I think are worthy of further research.

- **Computational tractability:** The primary goal of this project was to define a principled mathematical framework for value aggregation, which has been lacking in the literature, rather than to address computational tractability. Still, the computational times presented in Figure 5.3.1 indicate that, for similar applications, the computational requirements are well within acceptable limits for most practical uses. Some more research to explore how times evolve for more agents, values and actions would be an area of future improvement.
- **Value overlap:** The concept of value overlap is relatively new, with limited exploration in the literature (Karanik et al. 2024). Values that exhibit overlapping judgments, as illustrated in the toy example, may influence the aggregation process. Future work could explore how this overlap affects the aggregation outcome and develop methods that account for this phenomenon. Additionally, refining the definition of context and the structure of values remains an open research question.
- **Alternative metrics:** While the p-norm is well-established and supported by strong theoretical properties, other metrics could be considered for value aggregation. For example, distributional shifts, such as the Kullback-Leibler divergence (Kullback & Leibler 1951), could be applied to calculate VOA (and potentially AOA) as weights can be thought to come from a probability distribution. Ultimately, the p-norm is just one approach to aggregating value systems by seeking similar preferences over values, and exploring alternative metrics could enhance the robustness of the framework.
- **Value inference methodology:** One significant limitation of the existing literature on value inference, such as that proposed by Liscio et al. (Liscio et al. 2021, 2023, 2024), that they do not account for weighted preferences or shared interpretations of values. As argued in Section 2.2, this approach may not be appropriate for value aggregation. Future research should aim to develop a framework that better accommodates the value aggregation process already from the value inference setup.

As I embark on my PhD at the School of Science and Technology at City St George’s, University of London, I plan to further investigate the topics explored in this project. These future research efforts will build on the foundation laid by this work, addressing the open questions and limitations identified.

References

- Beauchamp, T. L. & Childress, J. F. (2019), *Principles of Biomedical Ethics*, 8th edn, Oxford University Press, New York.
- Brandt, F., Conitzer, V., Endriss, U., Lang, J. & Procaccia, A. D. (2016), *Handbook of computational social choice*, Cambridge University Press.
- Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S. & Yeung, K. (2021), Trustworthy ai, in ‘Reflections on Artificial Intelligence for Humanity’, Springer, pp. 13–39.
- Chisholm, R. M. (1963), ‘Supererogation and offence: A conceptual scheme for ethics’, *Ratio (Misc.)* **5**(1), 1.
- Commission, E. (2019), ‘Ethics guidelines for trustworthy ai’.
URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Dryzek, J. S. (2000), *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*, Oxford University Press, Oxford.
- European Union (2024), ‘Artificial intelligence act’.
URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>
- Gedeshi, I., Pachulia, M. & Poghosyan, G. (2017), ‘European values study 2017: Integrated dataset (evs 2017)’, GESIS Data Archive, Cologne. ZA7500 Data file Version 1.1.0.
URL: https://search.gesis.org/research_data/ZA7500?doi=10.4232/1.13560
- González-Pachón, J. & Romero, C. (2016), ‘Bentham, marx and rawls ethical principles: In search for a compromise’, *Omega* **62**, 47–51.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P. & Ditto, P. H. (2013), ‘Moral foundations theory: The pragmatic validity of moral pluralism’, *Advances in experimental social psychology* **47**, 55–130.
- Haidt, J. (2012), *The righteous mind: Why good people are divided by politics and religion*, Vintage.
- Karanik, M., Billhardt, H., Fernández, A. & Ossowski, S. (2024), On the relevance of value system structure for automated value-aligned decision-making, in ‘Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing’, SAC ’24, Association for Computing Machinery, New York, NY, USA, p. 679–686.
URL: <https://doi.org/10.1145/3605098.3636057>, doi = 10.1145/3605098.3636057
- Kullback, S. & Leibler, R. A. (1951), ‘On Information and Sufficiency’, *The Annals of Mathematical Statistics* **22**(1), 79 – 86.
URL: <https://doi.org/10.1214/aoms/1177729694>
- Lera-Leri, R., Bistaffa, F., Serramia, M., Lopez-Sanchez, M. & Rodriguez-Aguilar, J. (2022), Towards pluralistic value alignment: Aggregating value systems through lp-regression, in ‘Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems’, AAMAS ’22, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 780–788.
- Lera-Leri, R. X., Liscio, E., Bistaffa, F., Jonker, C. M., Lopez-Sanchez, M., Murukannaiah, P. K., Rodriguez-Aguilar, J. A. & Salas-Molina, F. (2024), ‘Aggregating value systems for decision support’, *Knowledge-Based Systems* **287**, 111453.
- Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R. I., Jonker, C. M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A. & Murukannaiah, P. K. (2023), Value inference in sociotechnical systems, in

- ‘Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems’, AAMAS ’23, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 1774–1780.
- Liscio, E., Siebert, L. C., Jonker, C. M. & Murukannaiah, P. K. (2024), ‘Value preferences estimation and disambiguation in hybrid participatory systems’.
URL: <https://arxiv.org/abs/2402.16751>
- Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., Mouter, N. & Murukannaiah, P. K. (2021), Axies: Identifying and evaluating context-specific values, *in* ‘Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems’, AAMAS ’21, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 799–808.
- Montes, N. & Sierra, C. (2021), Value-guided synthesis of parametric normative systems, *in* ‘Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems’, AAMAS ’21, IFAAMAS, Richland, SC, p. 907–915.
- Newberry, T. & Ord, T. (2021), ‘The parliamentary approach to moral uncertainty’, *Future of Humanity* .
- Osman, N. & d’Inverno, M. (2024), A computational framework of human values, *in* ‘Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems’, AAMAS ’24, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 1531–1539.
- Ross, W. D. (1930), *The Right and the Good*, Oxford University Press, Oxford.
- Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Penguin.
- Schwartz, S. H. (2012), ‘An overview of the schwartz theory of basic values’, *Online Readings in Psychology and Culture* **2**(1).
- Serramia, M., Lopez-Sanchez, M., Moretti, S. & Rodriguez-Aguilar, J. A. (2023c), ‘Building rankings encompassing multiple criteria to support qualitative decision-making’, *Information Sciences* **631**, 288–304.
URL: <https://doi.org/10.1016/j.ins.2023.02.063>
- Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A. & Moretti, S. (2024), Value alignment in participatory budgeting, *in* ‘Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems’, AAMAS ’24, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 1692–1700. Available here.
- Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Rodriguez, M., Wooldridge, M., Morales, J. & Ansotegui, C. (2018), Moral values in norm decision making, *in* ‘Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems’, AAMAS ’18, IFAAMAS, Richland, SC, p. 1294–1302.
- Serramia, M., Lopez-Sanchez, M. & Rodríguez-Aguilar, J. A. (2020), A qualitative approach to composing value-aligned norm systems, *in* ‘Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS ’20)’, IFAAMAS, Auckland, New Zealand, pp. 1233–1241.
- Serramia, M., Rodriguez-Soto, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Bistaffa, F., Boddington, P., Wooldridge, M. & Ansotegui, C. (2023b), ‘Encoding ethics to compute value-aligned norms’, *Minds and Machines* **33**, 761–790.

- Serramia, M., Seymour, W., Criado, N. & Luck, M. (2023), Predicting privacy preferences for smart devices as norms, *in* ‘Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems’, AAMAS ’23, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 2262–2270.
- Siebert, L. C., Liscio, E., Murukannaiah, P. K., Kaptein, L., Spruit, S. L., van den Hoven, J. & Jonker, C. M. (2022), Estimating value preferences in a hybrid participatory system, *in* ‘HHAI2022: Augmenting Human Intellect’, IOS Press, Amsterdam, the Netherlands, pp. 114–127.
- Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J. & Perello-Moragues, A. (2019), Value alignment: a formal approach, *in* ‘Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS’, IFAAMAS, Montreal, Canada.
- Soto, M. R., Serramia, M., López-Sánchez, M. & Rodríguez-Aguilar, J. A. (2022), ‘Instilling moral value alignment by means of multi-objective reinforcement learning’, *Ethics and Information Technology* **24**.
URL: <https://doi.org/10.1007/s10676-022-09635-0>
- Study, E. V. (2020), ‘European values study longitudinal data file 1981-2008 (evs 1981-2008)’.
URL: <https://europeanvaluesstudy.eu/>