# Finding our common moral values: Guidelines for value system aggregation

Anonymous Author(s)
Submission Id: 11

## ABSTRACT

The research community has already produced a breadth of approaches to resolve several value alignment problems. However, in the pursuit of value alignment, we usually need to know which values we want our AI to align with. This problem, called value inference, has caught some attention lately with many approaches to detect which moral values are relevant in a context, or to build a model (called value system) representing the values and priorities of an individual. However, another important task in value inference is that of value system aggregation. This consists in aggregating the moral value models of several individuals to obtain one representing everybody. So far, only one value system aggregation method has been proposed. In this paper, we discuss why research in value system aggregation is paramount and the possible avenues to implement value system aggregation depending on the value alignment problem at hand.

## CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; **Multi-agent systems**.

## KEYWORDS

Ethics, Value alignment, Aggregation, Participatory Budgets

**ACM Reference Format:**
Anonymous Author(s). 2025. Finding our common moral values: Guidelines for value system aggregation. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

## 1 INTRODUCTION

As Artificial Intelligence becomes commonplace, there are rising concerns about the ethical implications it has and will continue to bring about. As such, there has been increasing interest in value alignment [52], which is the task of ensuring that AI aligns with the moral values we want to uphold. In decision-making, one possible approach to ensuring value alignment is to consider a model of the moral values to guide the decision-making process [61]. When looking at how humans make decisions, we see that they usually consider several moral values. For example, we might decide to eat lettuce instead of steak because it is healthier, more sustainable, and cheaper. However, not all moral values may be relevant in a given context. For instance, when deciding what to eat, the value of benevolence might rarely be relevant. In addition to considering multiple values, we usually have preferences among them. For example, if the region we inhabit produces asparagus, we might prefer to eat asparagus over lettuce as they are more sustainable in this case, even though it might be more expensive, signalling that sustainability is preferred over economy. Thus, researchers in values and AI usually consider value systems [43, 60, 63] which, despite lacking a universal definition, commonly exhibit the aforementioned traits (multiple values and preferences among them). Value systems are thus models that can guide many value alignment approaches, so it is paramount to obtain them. However, finding the value system with which we want AI to align is a challenging problem.

Although the processes for obtaining value systems still require more detailed research, one proposal [38] approaches this problem in three steps, namely value identification, value system estimation, and value system aggregation. As previously discussed, different contexts may have different relevant values. For instance, in the context of alimentation, health and sustainability are relevant values whereas benevolence might not be relevant. The literature contains several universal sets of values, for example, Schwartz's theory of basic values [56], Hofstede's cultural dimensions [25], or those of the Moral Foundations theory [24]. Despite these sets serving as a good universal set of values, for value alignment applications it is better to consider a set of values specific to the context at hand. As such, the first step of value inference is **value identification**, which aims at finding the relevant values in a context. Value identification is usually approached by detecting the relevant values in context-specific texts, be it automatically [67] or semi-automatically [39]. The next step proposed by Liscio et al. is **value system estimation**, that is, to obtain the preferences of individuals over the previously detected relevant values. This again, can be achieved with text-based approaches [62]. Finally, and most relevant to the purpose of this paper, is **value system aggregation**, which aims at aggregating all the individual value systems of the previous step into a single system that represents everyone.

Crucially, while there has been significant research in both value system identification and value system estimation (see [38] for more details), value system aggregation has been mostly overlooked. In line with Conitzer et al. [9], we believe that Social Choice techniques can serve as a basis for this process. However, as we will discuss in later sections, preference aggregation techniques are not always suitable for performing value system aggregation, therefore we need to define new aggregation methods that can deal with the complexities brought by moral values. So far, only one value system aggregation approach has been proposed, that of Lera Leri et al. [35]. However, this approach is not without faults. While the authors aimed at building a value system aggregation method that is computationally feasible, they do not study its formal properties.

With this in mind, this paper aims to explore ideas for possible paths to value system aggregation, categorising several approaches

*Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.* 

we may wish to develop and providing illustrative use cases where they would be appropriate. In more detail, the contributions of this paper are:

- A formal characterization of value system aggregation as a distinct problem from preference aggregation.
- A taxonomy of value system aggregation approaches based on four key dimensions: judgment structure, ethical paradigms of aggregation, generality, and deliberation.
- An exploration of practical use cases illustrating how different aggregation approaches can address value alignment challenges in participatory policymaking.

The paper is structured as follows. In Section 2, we motivate value system aggregation as a useful tool in policymaking. Section 3 introduces necessary background on moral values and value systems. Section 4 discusses the differences between value system aggregation and preference aggregation. Sections 5, 6, 7, and 8 discuss possible avenues to define novel value system aggregation methods. Finally, Section 9 presents the conclusions and outlines directions for future work.

## 2 MOTIVATION: PARTICIPATORY BUDGETS

In this section, we motivate the relevance of value system aggregation for policymaking. As outlined by a European Commission technical report [55], knowing the moral values of citizens is important to understand them, co-create policy, and communicate it. In this discussion, we focus on how value systems can be used to inform policy creation.

Most democracies rely on a model of representative democracy, in which citizens elect their representatives in elections (usually every four years) and minimally participate in policymaking between them. However, representative democracy is facing a crisis, with citizens getting increasingly disengaged in elections. As shown by [19], although citizens remain interested in politics and want to make an impact, electoral participation has steadily declined since the 1980s worldwide. Participatory democracy serves as an alternative to representative democracy allowing citizens to get involved in day-to-day policymaking. Although there are many forms of participatory methods, like parliamentary petitioning [15, 48], we focus on participatory budgets.

Participatory budgeting allows citizens to propose and decide how to spend a government pre-defined budget. It is one of the most widely used participatory democracy tools. According to the Participatory Budget World Atlas [16], there were more than 10,000 participatory budgeting processes worldwide in 2021, including cities like New York [11], Santiago de Chile [14], Kakogawa [30], Madrid [41], Barcelona [3], Paris [46], or Cape Town [45]. Different PB processes are organised in various ways but they usually have three main phases: (1) Citizens propose ways to spend the budget; (2) Citizens vote on their preferred proposals; (3) Proposals are accepted or rejected depending on their citizen support.

Despite their widespread adoption, participatory budgets are not without faults. First, governments usually allocate relatively small budgets to participatory budget processes while the vast majority of public funds is managed without citizen involvement. For example, New York's participatory budget amounted to $30M in 2023, while the city's total municipal budget was $106 billion [12]. Second, and related to this paper, participatory budgets suffer from low democratic quality. Typically, participatory budgets suffer from low participation and unbalanced demographic representation. For example, Paris's participatory budget process, which is considered one of the most successful, had 137,622 participants in 2023 [47], which is around 6.5% of the city's population (2.1 million citizens). Furthermore, some studies suggest that participatory budget participants are usually from a wealthier background [53] (although, admittedly more studies on the demographics of participatory budget participants are needed). This setting is not fair to the immense majority of citizens who do not participate but contribute through taxes and are affected by the participatory budget outcomes. In contrast, an aggregated value system representing society can be used to correct the bias in participant-based votes. Serramia et al. [59] showed that by considering both the votes of the participants and a value system representing the society, participatory budgets could produce a solution that largely satisfied both participants and non-participants. In other words, using value systems we can find a participatory budget allocation that is almost optimal for the participant's priorities while compensating the possible biases introduced by them.

## 3 BACKGROUND: VALUE SYSTEMS

In this section, we discuss the necessary concepts for understanding values and introducing what value system aggregation is. First, we look into context.

The framework of value systems generally considers values as context-specific criteria. In applied ethics [4, 51], the scope of reasoning with values is often constrained by defining a *context* in which arguments acquire specific meanings and hold. This assumption is prevalent in the value systems literature [39, 44, 61], which prioritizes practical applications such as context-dependent decision-making. However, this raises the question of how contexts relate to one another and how they are defined—an issue explored in some empirical work [37]. Conversely, other frameworks analyse value systems without a contextual lens, such as Hofstede's cultural values in sociology [25] and the Moral Foundations Theory in psychology [24]. These approaches have also influenced recent research in value classification [37] and value overlapping [31]. Regardless of whether value systems are considered context-dependent, the challenge of aggregating different inputs remains a crucial question. In practice, we can represent contexts using any logical language; to be general, we do not restrict to one. Hereafter, we suppose a (possibly infinite) set of contexts, denoted $Con$, expressed as formulas in a given logical language.

A *value* is a deeply held belief that guides decision-making by defining what is important and desirable [56]. In essence, values allow us to discern which actions are good or bad. Therefore, we formally characterize values as utility functions that assess the desirability of actions. Following frameworks in applied ethics [8], given a set of possible actions $A$ and a set of possible contexts $Con$, each value $v \in V$ is defined by two judgment functions:

$$v^+, v^- : A \times Con \to [-1, 1] \tag{1}$$

The function $v^+(a, c)$ evaluates performing action $a \in A$ in the context $c \in Con$, while $v^-(a, c)$ assesses its non-performance. For

example, in the context of witnessing an accident on the road, the value of benevolence judges positively the performance of the action "help" and negatively its non-performance. To ensure coherence in moral evaluations and prevent a value from simultaneously assigning both positive and negative judgments to the same action [61], we impose the following constraint:

$$v^+(a,c) \cdot v^-(a,c) \leq 0, \quad \forall a \in A, c \in Con. \tag{2}$$

A *value system* is composed of a set of moral values and their relative importance. Formally, we define a value system as follows:

*Definition 3.1.* A *value system* is a tuple $VS = \langle V, \succcurlyeq \rangle$, where $V$ is a set of moral values and $\succcurlyeq$ is a preference relation over these values.

This general definition is widely referenced in the literature [1, 63]. Some authors [20, 35] extend this framework by allowing values to have agent-specific interpretations. In this approach, while the set of value labels $V$ remains shared within a given context, the structure of each value —defined by its judgment functions $(v^+, v^-)$— varies across agents. We explore this distinction further in Section 5. Additionally, there is variation in how preferences between values are represented, with different levels of input constraints. Examples include pairwise preferences [36], complete rankings [58], and normalized weighted preferences [26].

Given a set of agents $Ag$, the process of *value system aggregation* consists of combining individual value systems into a single consensus value system that can represent the group of agents.

*Definition 3.2.* Given $Con$ a set of contexts, $A$ a set of actions, and $\mathcal{VS}$ the set of all possible value systems with values defined over $A \times Con$, a *value system aggregation function* $F$ is a mapping:

$$F : \mathcal{VS}^n \rightarrow \mathcal{VS}, \tag{3}$$

where $n = |Ag|$ represents the number of input value systems.

## 4 PREFERENCE VS. VALUE SYSTEM AGGREGATION

Now that we understand what value systems are, in this section we address why value system aggregation is needed and how it is different from preference aggregation.

Researchers in Social Choice have long studied the problem of aggregating individual inputs —such as votes, preferences, or judgments— into a collective outcome. Since value systems encode a preference structure over a set of relevant values within a given context, classic preference aggregation rules could, in principle, be applied directly, leveraging their well-studied mathematical properties. For instance, if the input value systems consist of complete orderings of values, standard aggregation methods such as the Kemeny rule [32] can be employed to derive a consensus ranking. Beyond Arrow's foundational framework for preference aggregation [2], the value alignment field can also benefit from research like Sen's extension [57], which accommodates cardinal preferences and enables interpersonal comparisons, enriching the possibilities for value system aggregation.

However, preference aggregation alone may not fully capture the structure of value systems. If values were merely options to be ranked, traditional preference aggregation would suffice. Yet, values are not just ranked alternatives; they are composed of judgment functions that define their meaning and thus their role in decision-making. While a naive aggregation approach might approximate a consensus, it risks overlooking the complexity of value systems. One key challenge arises if we allow agents to interpret moral values differently, as in this case each agent considers different judgement functions for the values. Lera-Leri et al. [35] address this problem by proposing a method that combines not only value preferences but also individual interpretations, distinguishing value system aggregation from traditional social choice mechanisms. See more in section 5.

By considering the structure of value systems—encompassing both shared and individual judgment functions—we can account for ethical concerns like plurality. This connects to the broader paradigm of Ethics by Design, which emphasizes embedding ethical principles into system development from the outset to proactively address ethical issues rather than reacting to them later. Discussed prominently in the context of autonomous systems by Dignum [17] and later formalized as a research and development framework by the European Commission [54], this approach seeks to anticipate negative consequences and safeguard fairness and inclusivity. For instance, it would favour integrating bias mitigation into an AI recruitment system from the start, rather than correcting discrimination afterward. In value system aggregation, this perspective is particularly relevant when choosing an ethical paradigm—such as fairness-driven approaches (e.g., Rawlsian fairness)—to ensure minority voices are not drowned out by majority preferences. Embedding such considerations into the design of aggregation methods is crucial to **selecting an appropriate ethical paradigm** as part of a holistic approach to aggregation—be it utilitarian, Rawlsian, or an intermediate approach—to balance overall societal satisfaction with adequate representation of minority perspectives. See more in section 6.

Another crucial distinction is that value system aggregation would often be performed with **a specific application in mind**, such as producing a value system representing citizens to be used in participatory budgets. This process involves not only determining the relative importance of values but also understanding how their judgment functions shape decisions. Unlike preference aggregation, which combines several preferences into one, value system aggregation must combine the value preferences, the value judgments (if the input ones are different) and all this ensuring the value system produced is adequate for the application in mind. For example, when finding a value system for a participatory budget typical preference aggregation methods will indeed produce an aggregated value preference. However, a value system aggregation method should also consider how the preferences will stir the participatory budget afterwards. For example, extending the Kemeny rule, a value systems aggregation could minimise the distance between the decisions made by the aggregated value system and those made by the input ones. This added layer of complexity arises because preferences are initially defined over values, yet the ultimate goal is to use them to guide action choices in a value-aligned way. See more in section 7.

Besides the way we formalize aggregation mathematically, there are contexts— such as reconciling cultural differences in public governance—where a computational approach alone may not be

suitable. This calls for considering alternative methods, such as **deliberation-based processes** or hybrid approaches, where a collective value system emerges through group reasoning rather than preconceived mathematical formulations. See more in Section 8.

## 5 PERSONAL VS UNIVERSAL JUDGEMENTS

A fundamental consideration when designing a value system aggregation approach is whether all agents share a common understanding of the relevant values. If interpretations vary, the aggregation process must reconcile these differences by establishing a consensus interpretation for each value along with the collective preferences. This ensures that the output maintains the same structure as the inputs. Conversely, if all agents align in their interpretation of values, the aggregation process is simplified in this regard, as the structure of values remains consistent throughout.

### 5.1 Personal judgements

Approaches based on individual value interpretations recognize that agents may prioritize the same value while differing in how they define it. For example, two agents may think security is the most important value, but they might understand security differently, with one advocating for widespread firearm ownership to protect oneself while the other favours strict gun control to minimise gun violence.

Despite the high fidelity of this approach to an agent's stated motivations, aggregating value preferences and interpretations into a single consensus value system presents a theoretical challenge. If agents assign the highest priority to a value but interpret it in opposing ways —such as in the security example above— then the consensus output will still rank this value as most important, yet the aggregated interpretation may fail to align with any of those of the stakeholders. This underscores a key issue: value preferences among agents do not operate over a shared set of moral values, as they are inherently shaped by individual interpretations. Consequently, even if an agent's preferences align with the consensus, they may still feel unrepresented if their understanding of the values differs significantly from the aggregated interpretation.

Use Case. *Aggregation approaches considering personal judgements are useful in cases where the agents have small differences in their understanding of how values judge actions but do not hold opposite views. For example, imagine a city with a lot of crime where policymakers want to design policy to stop it. They will guide their decisions using the value system of the population, security is a relevant value in this case. They know through some preliminary surveys that citizens have similar views on this and other relevant values, then they can use personal judgements in the aggregation to find the population value system.*

### 5.2 Universal judgements

Approaches based on a universal value understandings constrain individual interpretations of values, requiring each agent to express their value system preferences using a predefined set of inferred value judgements. This approach can be based on the assumption that values assess actions according to objective criteria that remain consistent across agents. For instance, in the context of route selection for driving, the value of sustainability evaluates choices based on fuel efficiency, while the value of security assesses them according to the risks associated with the selected roads [26]. These approaches showcase less flexibility to capture the nuances of individual justifications and can be seen as reducing ethical dilemmas to multi-criteria decision making.

Use Case. *Aggregation approaches considering universal judgements are useful in cases where agents have large differences or opposing views on how values judge some actions. Continuing on the design of policy to prevent crime, if some agents think the value of security judges positively the action of carrying guns, while others think this action is judged negatively, then we should use universal judgements instead of personal ones.*

### 5.3 Summary and possible alternatives

| | Universal judgements | Individual judgements |
|---|---|---|
| **Representation** | May not capture all perspectives. | Custom interpretations; diverse perspectives. |
| **Comparability** | Preferences directly comparable. | Preferences not directly comparable. |
| **Value Identifi.** | Hard. Predefined labels and shared judgments. | Easy. Labels without shared judgments. |
| **Value Estimation** | Easy. Estimate preferences only. | Hard. Estimate preferences and judgments. |
| **Consensus Output** | Clear, interpretable consensus value system. | Value judgements may lose meaning. |

**Table 1: Comparison of Universal and Individual Value Approaches.**

Table 1 provides a summary of the benefits and drawbacks for each approach discussed above, however note we can also find alternative approaches.

One potential compromise to address the challenges of aggregating value systems with differing interpretations involves relaxing the completeness requirement of the consensus value system. To avoid aggregating conflicting interpretations of a value, a similarity threshold can be introduced based on a spread measure (e.g., variance). This ensures that only values with reasonably close interpretations are aggregated into the consensus, contributing to the preference relation. Values with interpretations that exceed the threshold are deemed non-aggregable, resulting in an incomplete value system that represents stakeholders only for reconcilable values. Another compromise approach, as suggested in [38], is to aim for multiple consensuses when individuals cluster around distinct interpretations, rather than forcing a single consensus value system that may not adequately represent any group.

From the perspective of the universal approach, one potential compromise to address its limited representation of diverse perspectives is to focus on extensive value identification. The rationale is that, when different clustered interpretations of a value exist (as in the earlier example of security), effective value identification should distinguish as many values as there are clusters, rather than averaging interpretations or ignoring some. Different interpretations

capture distinct nuances of values. For instance, instead of a single value like "security," we could identify "public security" and "private security" to account for previously conflated perspectives. Through exhaustive value identification, no relevant perspective would be overlooked, and preferences would comprehensively encompass all significant reasoning behind actions. This compromise could exemplify how value system aggregation should inform value estimation, aligning with the observation that the interplay between value inference steps remains an under-explored but essential research direction in the field [38].

# 6 AGGREGATION ETHICAL PARADIGMS

In addition to the differences between value judgment frameworks, other significant aspects of aggregation warrant consideration, such as aggregation ethical paradigms. Distance-based preference aggregation is a widely used aggregation framework that involves defining a distance between aggreganda and selecting a consensus aggregandum that minimises the total distance to all inputs. Thus, given candidates $C$, a set of agents $1, \ldots, n$, and preference order $\succ_i$ for each agent $i$, and a distance function $d$ between preference orders, the essence of distance-based aggregation is to find:

$$\succeq_{agg} = \underset{\succeq \subseteq Con \times Con}{\arg\min} \sum_{i=1}^{n} d(\succ, \succ_i) \tag{4}$$

Distance-based aggregation rules are versatile, as they can be applied to any kind of input for which a distance metric is defined, such as preference orders or welfare functions [5]. The Kemeny rule exemplifies a distance-based aggregation method, as it minimizes Kendall's tau distance [33], which measures the number of pairwise disagreements between two preference orders.

One important issue to bear in mind in value system aggregation is the ethics of the aggregation process itself. Following the principles of Ethics by Design, it is paramount to ensure, at all stages of the production and use of a value system, that it adheres to our ethical goals. For example, we could produce a value system that is as similar to all individual value systems as possible, or we could maximise fairness towards outliers and increase their impact on the outcome. In distance-based preference aggregation, the work of Gonzalez-Pachón et al. [22] addresses aggregation ethical principles by considering a p-metric distance function, which we can generalise to value system aggregation. Thus, given agents $1, \ldots, n$, a value system $vs_i$ for each agent $i$, a distance function $d$ between value systems, and let $\mathcal{VS}$ be the set of all possible value systems, we can then define distance-based value system aggregation incorporating the ethical parameter $p \in [1, +\infty)$ as follows:

$$vs_{agg} = \underset{vs \in \mathcal{VS}}{\arg\min} \left( \sum_{i=1}^{n} d(vs, vs_i)^p \right)^{\frac{1}{p}} \tag{5}$$

Note that this equation defines a family of value system aggregation functions that depend both on the value system distance function $d$ and the parameter $p$. The $p$ parameter allows for different considerations of the trade-offs between overall utility, individual fairness, and the influence of extreme positions, making the choice of $p$ a critical design decision that depends on the context of application. Next, we discuss some particularly interesting values of $p$ and study their usefulness for value system aggregation.

## 6.1 Utilitarian Aggregation ($p = 1$)

The case $p = 1$ corresponds to the minimisation of the sum of raw distances. This approach treats all distances equally, making it a utilitarian method that minimizes the overall discrepancy among agents with respect to the consensus value system. Mathematically, the problem is expressed as:

$$vs_{ut} = \underset{vs \in \mathcal{VS}}{\arg\min} \sum_{i=1}^{n} d(vs, vs_i),$$

where $n$ is the number of agents, $vs_i$ represents the value system of agent $i$, $\mathcal{VS}$ is the set of all possible value systems, and $d$ is the value system distance function. This method is particularly effective in large-scale aggregation scenarios, where the aim is to capture the general tendency of thousands or millions of agents' value systems.

However, if no further constraint is imposed on the output beyond distance minimization, it invariably results in a median solution, meaning the consensus value system is basically unaffected by outliers. Thus, utilitarian value system aggregation can lead to under-representation in pluralistic contexts.

Use Case. *Utilitarian approaches may be suitable when the population and their opinions on values are fairly homogenous as there is no need to give special attention to outliers. For example, imagine that a participatory budget succeeds in mobilising minorities, then even if we use a value system to compensate for non-participants, we can use a utilitarian aggregation as minorities might already be overrepresented in the participant base. A utilitarian approach is also useful if the decision-maker wants to under-represent outliers. For example, a participatory budget organiser might decide to use a utilitarian aggregation if the non-participant base contains a small fraction of the population whose only goal is to undermine the participatory budget and the democratic quality of the city.*

## 6.2 Rawlsian Fairness ($p = \infty$)

The case of $p = \infty$, corresponds to a Rawlsian approach to aggregation. Rawls principle of justice as fairness [50] consists on minimising the maximum distance between individual and aggregated value systems to ensure fairness for disadvantaged stakeholders. Hence:

$$vs_{fair} = \underset{vs \in \mathcal{VS}}{\arg\min} \max_{i=1}^{n} d(vs, vs_i).$$

This approach is particularly relevant in contexts of marginalized communities or highly unequal input distributions. However, by focusing exclusively on extreme positions, it risks producing a consensus that diverges from majority preferences, potentially reducing its practicality when overall group alignment is a primary objective.

Use Case. *As we have discussed fairness-based aggregations are useful when we want outliers to have increased impact on the result. This is useful in cases were minority representation is relevant. For*

| | Utilitarian Aggregation ($p = 1$) | Rawlsian Fairness ($p = \infty$) | Intermediate Approaches ($p > 1$) |
|---|---|---|---|
| **Risk of Bias** | Majority domination; ignores minority positions | Minority over-representation; can diverge from majority preferences | Potential for balanced representation |
| **Suitability with low dispersion input** | Best. The result will be representative | Good. There are no major outliers, so it is similar to utilitarian aggregation. | Good. The solution will be similar to the utilitarian. |
| **Suitability with high dispersion input** | Bad. Several agents will not be represented | Best. The solutions is the fairest to outliers. | Good. Compromise between utilitarianism and fairness. |
| **Applicability** | Readily applicable | Readily applicable | Finding a suitable $p$ might be hard |

**Table 2: Comparison of Aggregation Paradigms in Value System Aggregation. Intermediate approaches represent a family of approaches and thus the optimal $p$ for each situation remains a decision to be made based on the principles highlighted.**

*example, in participatory budgets some proposals may be minority-specific, and these will have little chance of getting funded unless we help by over-representing the minority in the value system. For example, a proposal related to accessibility (e.g. converting stairs to ramps) only benefits people that have accessibility requirements while it is irrelevant for the rest of citizens. However, projects of this sort greatly impact the lives of people with accessibility requirements. It is possible that even if this minority mobilises to vote, the project will not get enough votes to be funded. Hence, by over-representing this minority in the value system we give more chances to this kind of proposals of getting funded.*

### 6.3 Intermediate approaches ($p > 1$)

When $p > 1$, the aggregation process gives greater weight to larger distances, thereby increasing the influence of outliers as $p$ increases without fully determining the aggregation. For instance, in the case of $p = 2$, the problem minimizes the squared distances (which yields the mean in the unconstrained optimization problem):

$$vs_{p=2} = \arg\min_{vs \in \mathcal{VS}} \sqrt{\sum_{i=1}^{n} d(vs, vs_i)^2}.$$

The choice of $p$ directly shapes the aggregation outcome, marking a continuous trade-off between equality of all agents ($p = 1$) and fairness towards the most outlying agents ($p = \infty$). Lower values of $p$ favour equal impact of all agents while higher values shift towards a fairness-driven approach, amplifying the influence of those most distant from the consensus. This trade-off is particularly relevant in pluralistic societies, where ensuring adequate representation while maintaining a coherent consensus is essential.

Understanding this spectrum is crucial for value system aggregation, as different scenarios may call for different ethical priorities, balancing collective agreement with the need to protect minority perspectives.

Use Case. *Intermediate approaches will be useful in cases when we want to maintain an equal representation of all agents but we have uneven participation. For example, as previously discussed high-income citizens tend to participate more than low-income ones in participatory budgets. Thus, if we use a utilitarian aggregation approach high-income individuals will be overrepresented, otherwise if we use a fairness-based aggregation we will over-represent low-income citizens. If we consider income is not a trait worth giving special importance, then we know there is some value of $p$ that is able to give enough increased importance to low-income citizens so that the bias in the participant base is compensated but not over-compensated. Of course, in real cases there are multiple factors to consider (the population will probably have widely differing income, age, there might be many different minorities...) so deciding on the value of $p$ is not straightforward. In the end, the decision-maker has to weigh all these factors and decide $p$ accordingly.*

To conclude this section, Table 2 summarises the benefits and drawbacks of each approach.

## 7 GENERAL VS TAILORED AGGREGATION

Another important point to decide when designing or choosing value system aggregation methods is whether they are tailored to a specific application or if they are general value system aggregation methods. As previously mentioned in the introduction, the only value system aggregation method introduced so far [35] was not designed with any particular application in mind. Furthermore, the authors do not study its social choice properties either, hence it is difficult to see what applications it could be useful for. Admittedly, and as argued by Lera Leri et al. [36], their lp-regression method does satisfy some classic social choice properties in very special cases, as supported by the research in [21]. However, as discussed in Sec. 4 value system aggregation defines a new paradigm that generalises typical preference aggregation, so in some cases classic social choice properties may apply, but in most cases we will need to generalise them or formalise new moral value-focused properties. Hence, we will refer to aggregation methods like Lera Leri et al.'s [35] as general aggregation approaches.

Use Case. *General approaches might be useful, for example, for a small government that wants a model of its citizens' value system to use it for several tasks such as informing policy-making, policy communication or participatory budgets and is seeking a quick and cheap off-the-shelf solution. Arguably, even though general approaches are not optimal, if a government does not have a lot of resources it is better to use a general value system aggregation approach than to not consider the citizens' values at all.*

Alternatively, we can design a value system aggregation approach with the aim of satisfying some moral value-focused properties or with some application in mind. We call these, tailored value

system aggregation approaches. Table 3 outlines the main benefits and drawbacks of both types of approaches. In essence, the more tailored an aggregation approach is, the more money, research and probably computational resources will be needed to make it a reality. As the usefulness of tailored approaches has not been discussed previously in the values community, we provide a possible avenue for producing a tailored value system aggregation approach.

| | General Approaches | Tailored Approaches |
|---|---|---|
| **Availability** | Readily available (e.g. [35]). | Needs more research (not existent yet). |
| **Resources** | Minimal economic and computational resources needed. | Economic resources will be needed, there is no evidence of computational tractability. |
| **Optimality** | Produce an approximate value system aggregation for most applications | Decision-makers will know the value system they use is optimal for the task at hand. |

**Table 3: Comparison of general and tailored value system aggregation approaches.**

### 7.1 Example tailored value system aggregation: Minimal decision divergence

A key motivation for tailored value system aggregation methods is the need to align the aggregated value system with specific decision-making requirements. While value systems are particularly useful for ensuring interpretability and understanding the rationale behind choices, values themselves represent relatively general criteria. In contexts where the number of available actions is limited, it may therefore be advantageous to reason about the aggregated value system at the decision level, ensuring that the actions it prescribes align closely with those of the input individual value systems. This motivates the design of an aggregation method that aims to minimize the deviation between the actions prescribed by the aggregated value system and those supported by the individual agents, while preserving the underlying value-based reasoning. We refer to this principle as **Minimal Decision Divergence**.

As discussed in Section 6, one approach to comparing value systems is through their judgments of actions. Value systems inherently define a preference relation over actions, allowing for a measure of dissimilarity based on their decision-making implications. The relative importance assigned to different actions by an individual's value system reflects their choices in decision-making contexts, capturing how actions align with their underlying values. We can define a **decision divergence distance** as a function measuring the difference between two value systems based on their induced preferences over actions. Depending on how the utility over actions is structured—whether through pairwise comparisons, complete orderings, or cardinal utilities—this distance can be formulated in multiple ways.

Having established a decision divergence distance between two value systems, we can now determine the candidate value system that minimizes the distances to all input value systems, which we refer to as the consensus. The principle of **minimal decision divergence** aims to produce a consensus value system whose induced preferences over actions deviate as little as possible from those of the input agents. Rather than aggregating preferences or values independently, this approach ensures that the collective value system preserves decision-making coherence with the original inputs, prioritizing consistency in action recommendations. Since aggregation need not follow a simple linear sum (as detailed in Section 6), alternative formulations can be used to reduce large individual disagreements, balancing fairness and representation.

Unlike general aggregation approaches, which aim for broad applicability, an aggregation based on minimal decision divergence is inherently a *tailored* method. It is particularly suited for applications where preserving action alignment is critical, such as decision-support systems or ethical AI frameworks that require transparent justifications for recommended actions. Minimising the decision divergence is a novel desirable value-related property, therefore a method satisfying this property will have a more focused applicability than general aggregation methods for which value-related properties have not been studied.

Use Case. *A tailored value system aggregation based on the minimal decision divergence principle would be useful in a participatory budget where citizens have different ideologies and it is primordial that we try to find the best consensus to avoid citizen disaffection with the process. Importantly, since the complexity of the aggregation depends on the number of actions/decisions judged by the values, this method would be particularly suited for cases in which the number of possible actions/decisions is small. However, in cases where we have a larger number of actions/decisions, the aggregation can be performed based on a sample of them, though in this case we would only have an approximation to minimal decision divergence and not the optimal value system aggregation.*

## 8 ALTERNATIVES TO AGGREGATION

Thus far, we have examined various aspects of value system aggregation, including the role of individual judgments, the ethical principles guiding distance-based aggregation methods, and the distinction between tailored and general approaches. However, apart from formal aggregation, value systems can also be found with other techniques, for example, methodologies that explicitly incorporate structured deliberation and group reasoning.

This section explores *group decision-making* (GDM), a framework encompassing decision processes where multiple individuals contribute their judgments and preferences to form a collective outcome [7]. A key subset of GDM is *multi-expert decision-making* (MEDM), wherein a panel of experts provides structured input to refine and converge on a decision [27]. The MEDM framework could be particularly relevant to obtain a societal value system as it has been extensively studied across disciplines, including ethics-based AI governance [6], policy-making [29] and forecasting [64]. By leveraging insights like GDM and MEDM, we aim to compare two distinct paradigms: the **aggregation-based** methods discussed so far, and the **deliberation-based** methods.

Deliberation-based approaches prioritize iterative refinement through structured discourse, aiming to reach a collective decision through voluntary compromises rather than by selecting winners

and losers. Examples of this are the General Assemblies of the Occupy movement [23] or the citizen assemblies [10, 66]. These methods foster an implicit synergy, as participants engage in dialogue that may reveal overlooked arguments aligning with their underlying concerns. By facilitating open discussion, deliberation-based approaches encourage mutual understanding and consensus-building, in contrast to aggregation-based methods, which can be seen as more consensus-imposing.

While deliberative approaches offer flexibility and the potential for deeper consensus, they do not always produce consistent outcomes. Decisions depend on contextual factors and discussion dynamics rather than a fixed algorithmic process, making them difficult to replicate. This variability underscores the biases introduced by deliberation due to its reliance on participant interactions, group composition, and external influences. Furthermore, deliberative settings are not immune to dysfunctional group dynamics, such as *groupthink*, where the pressure for unanimity suppresses dissenting viewpoints [28], and *group polarization*, which can push participants toward more extreme positions rather than fostering balanced compromise [65]. Additionally, deliberation is inherently time-consuming, and requires active participation, and as argued in Section 2, it is hard to mobilise citizens. Table 4 summarises the benefits and drawbacks of each approach.

| | **Aggregation-Based Methods** | **Deliberation-Based Methods** |
|---|---|---|
| **Imposition** | Aggregation may impose an outcome. | Encourages voluntary compromises. |
| **Reliability** | Satisfies formal social choice properties | Prone to social biases (groupthink, polarization, dominant voices.) |
| **Consistent outcomes** | Yes. Same inputs yield same results. | No. Context and discussion affect results. |
| **Time Consuming** | Fast. One-shot decision-making. | Slow. Requires iterative refinement. |
| **Necessary input** | Requires individual value systems obtained beforehand | There is no required input beforehand. |
| **Participation** | No recruiting necessary | Recruiting citizens might be a challenge |

**Table 4: Comparison of aggregation-based and deliberation-based methods to obtain a societal value system.**

While aggregation-based and deliberation-based methods represent two distinct paradigms, many real-world decision-making processes incorporate elements of both. For instance, electoral systems often involve a public deliberation phase where priorities and policy discussions shape voter perspectives before a final aggregation of votes determines the outcome. In the context of multi-expert decision making, an example that bridges these approaches is the *Delphi method*, a structured forecasting technique first introduced in [13]. The method employs multiple rounds of questionnaires where expert opinions are collected, summarized, and redistributed anonymously, mitigating biases like groupthink and guiding experts toward consensus through iterative refinement. While the final decision relies on aggregated responses, the key feature of the Delphi method is its emphasis on structured synthesis rather than purely mathematical aggregation, underscoring its hybrid nature in decision-making frameworks. More broadly, evidence from social choice suggests that group deliberation, even when failing to produce full agreement, can foster a meta-agreement on the terms and concepts under discussion, simplifying subsequent aggregation [18, 34, 40, 42, 49]. In the context of value systems, this may encourage agents to converge toward shared value judgments, reducing the discrepancies arising from individual interpretations which as explained in Section 5 introduce complexity to the process.

These hybrid approaches illustrate that the process of obtaining a societal value system may benefit from integrating both aggregation mechanisms and structured deliberation, balancing reasonable mathematical properties with the depth of collective reasoning.

Use Case. *Value inference is a data intensive process, current approaches [38] require value-based text justifications of action decisions to produce individual value systems. Thus, a small government with limited budget may not be able to run the necessary surveys to collect data to perform value inference and aggregation. For example, if a government with limited resources wants a value system for participatory budgets, deliberative approaches represent a cheaper and easier to implement alternative to value system inference and aggregation. However, deliberative approaches depend on citizen participation, so they might suffer from the same bias and unequal representation of participatory budgets. Conversely, big data driven value inference and aggregation will perform better in this regard.*

## 9 CONCLUSIONS AND FUTURE WORK

Value alignment requires that we know what value system we want to align with. While the literature has begun to address the problem of obtaining value systems from society, it remains a largely open problem. In this paper, we have explored potential avenues for future research on the task of value system aggregation. In essence, the various properties and types of value system aggregation discussed in this paper are not mutually exclusive but together form a characterisation of possible aggregation methods, each with its different combined use cases. Note that, the only value system aggregation method currently available [35] is a personal judgement, general aggregation method which covers all ethical paradigms we considered (as it can be adapted through a parameter). However, as we have seen throughout the paper, there might be other use cases for which this approach is not useful. For example, imagine a small government wants to use a value system representing its citizens to compensate for the low participation of participatory budgets. The government already possesses survey data that can be used to create individual value systems for some of its citizens; however, the data is not detailed enough to account for personal judgements. Moreover, the government requires a value system specifically tailored to use in participatory budgetss. In this case, the government would benefit from a universal judgement, tailored aggregation method. In future work, we plan to develop alternative value system aggregation methods to cover these gaps in the literature. Our first step will be to study the minimal decision divergence property discussed in Section 7.1 as it is particularly useful in participatory budgets.

# REFERENCES


[1] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. 2017. Value Alignment or Misalignment – What Will Keep Systems Accountable?. In *Proceedings of the 2017 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, US, 15–21.

[2] Kenneth J Arrow. 1951-2012. *Social choice and individual values*. Vol. 12. Yale university press, New Haven, Connecticut.

[3] Decidim Barcelona. 2020. Pressupostos participatius 2020-2023. Accessed August 2023.

[4] Tom L Beauchamp and James F Childress. 2019. *Principles of Biomedical Ethics* (8th ed.). Oxford University Press, New York.

[5] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of computational social choice*. Cambridge University Press, Cambridge, UK.

[6] Emily Brown. 2020. Collaborative Frameworks for Ethical AI Governance. *Journal of AI Ethics* 5, 2 (2020), 123–145. https://doi.org/10.1007/s00146-020-00999-9

[7] CT Lawrence Butler and Amy Rothstein. 1991. *On conflict and consensus: A handbook on formal consensus decisionmaking*. Food Not Bombs, Cambridge, MA.

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

[9] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. 2024. Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. *arXiv preprint arXiv:2404.10271* (2024).

[10] Convention Citoyenne pour le Climat. 2020. Final Report. https://www.conventioncitoyennepourleclimat.fr/en/. Accessed: 2025-02-10.

[11] New York City Council. 2023. Participatory budgeting PBNYC. https://council.nyc.gov/pb/. Accessed Oct 2023.

[12] New York City Council. 2024. City budget. https://council.nyc.gov/budget/. Accessed 02/2024.

[13] Norman Dalkey and Olaf Helmer. 1963. An experimental application of the Delphi method to the use of experts. *Management Science* 9, 3 (1963), 458–467.

[14] Municipalidad de Santiago de Chile. 2023. Presupuestos Participativos Santiago. https://www.munistgo.cl/presupuestos/. Accessed Oct 2023.

[15] Plateforme des pétitions. 2024. Assemblée Nationale. https://petitions.assemblee-nationale.fr. Accessed 02/2024.

[16] Nelson Dias, Sahsil Enríquez, Rafaela Cardita, Simone Júlio, and Tatiane Serrano. 2021. *Participatory budgeting world atlas*. Epopeia and Oficina, Vila Ruiva - Portugal.

[17] Virginia Dignum. 2018. Ethics by Design: Necessity or Curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, US, 60–66. https://doi.org/10.1145/3278721.3278745

[18] John S. Dryzek and Christian List. 2003. Social Choice Theory and Deliberative Democracy: A Reconciliation. *British Journal of Political Science* 33, 1 (2003), 1–28.

[19] International Institute for Democracy and Electoral Assistance. 2023. Global State of Democracy Indices. https://www.idea.int/democracytracker/gsod-indices/. Accessed on February 2025.

[20] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[21] Jacinto González-Pachón and Carlos Romero. 2015. Properties underlying a preference aggregator based on satisficing logic. *International Transactions in Operational Research* 22, 2 (2015), 205–215. https://doi.org/10.1111/itor.12116 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/itor.12116

[22] Jacinto González-Pachón and Carlos Romero. 2016. Bentham, Marx and Rawls ethical principles: In search for a compromise. *Omega* 62 (2016), 47–51. https://doi.org/10.1016/j.omega.2015.08.008

[23] David Graeber. 2013. *The Democracy Project: A History, a Crisis, a Movement*. Spiegel & Grau, New York, US.

[24] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, Patricia Devine and Ashby Plant (Eds.). Advances in Experimental Social Psychology, Vol. 47. Academic Press, New york, US, 55–130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

[25] Geert Hofstede. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture* 2, 1 (2011), 8.

[26] Andrés Holgado-Sánchez, Javier Bajo, Holger Billhardt, Sascha Ossowski, and Joaquín Arias. 2024. Value Learning for Value-Aligned Route Choice Modeling via Inverse Reinforcement Learning. (June 2024). https://hal.science/hal-04627792 Submitted to VALE (VALUE ENGINEERING IN AI) track of the International Workshop on AI Value Engineering and AI Compliance Mechanisms (VECOMP 2024), affiliated with the 27th European Conference on Artificial Intelligence (ECAI 2024).

[27] Van-Nam Huynh and Yoshiteru Nakamori. 2005. Multi-Expert Decision-Making with Linguistic Information: A Probabilistic-Based Model. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 3 - Volume 03 (HICSS '05)*. IEEE Computer Society, USA, 91.3. https://doi.org/10.1109/HICSS.2005.448

[28] Irving L. Janis. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes* (2nd ed.). Houghton Mifflin, Boston.

[29] Michael Jones. 2015. Integrating Expert Opinions in Public Policy Development. *Policy Studies Review* 32, 1 (2015), 89–110. https://doi.org/10.1111/psr.12075

[30] Kakogawa. 2023. Make our Kakogawa. https://kakogawa.diycities.jp/. Accessed Oct 2023.

[31] Marcelo Karanik, Holger Billhardt, Alberto Fernández, and Sascha Ossowski. 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* (Avila, Spain) *(SAC '24)*. Association for Computing Machinery, New York, NY, USA, 679–686. https://doi.org/10.1145/3605098.3636057

[32] John G. Kemeny. 1959. Mathematics without Numbers. *Daedalus* 88, 4 (1959), 577–591. http://www.jstor.org/stable/20026529

[33] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (1938), 81–93. http://www.jstor.org/stable/2332226

[34] Jack Knight and James Johnson. 1994. Aggregation and Deliberation: On the Possibility of Democratic Legitimacy. *Political Theory* 22, 2 (1994), 277–296.

[35] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar. 2022. Towards Pluralistic Value Alignment: Aggregating Value Systems Through lp-Regression. In *Int. Conf. on Autonomous Agents and Multiagent Systems* (New Zealand) *(AAMAS '22)*. IFAAMAS, Richland, SC, 780–788.

[36] Roger X. Lera-Leri, Enrico Liscio, Filippo Bistaffa, Catholijn M. Jonker, Maite Lopez-Sanchez, Pradeep K. Murukannaiah, Juan A. Rodriguez-Aguilar, and Francisco Salas-Molina. 2024. Aggregating value systems for decision support. *Knowledge-Based Systems* 287 (2024), 111453. https://doi.org/10.1016/j.knosys.2024.111453

[37] Enrico Liscio, Alin E. Dondera, Andrei Geadău, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. Cross-Domain Classification of Moral Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2727–2745. https://doi.org/10.18653/v1/2022.findings-naacl.209

[38] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. 2023. Value Inference in Sociotechnical Systems. In *Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS '23)*. IFAAMAS, Richland, SC, 1774–1780.

[39] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. What values should an agent align with? *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 23. https://doi.org/10.1007/s10458-022-09550-0

[40] Christian List, Robert C. Luskin, James S. Fishkin, and Iain McLean. 2013. Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls. *The Journal of Politics* 75, 1 (2013), 80–95.

[41] Decide Madrid. 2022. Presupuestos participativos 2022. https://decide.madrid.es/presupuestos. Accessed August 2023.

[42] David Miller. 1992. Deliberative Democracy and Social Choice. *Political Studies* 40, S1 (1992), 54–67.

[43] Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. IFAAMAS, Richland, SC, 907–915.

[44] Pablo Noriega and Enric Plaza. 2024. On Autonomy, Governance, and Values: An AGV Approach to Value Engineering. In *Value Engineering in Artificial Intelligence*, Nardine Osman and Luc Steels (Eds.). Springer Nature Switzerland, Cham, 165–179.

[45] City of Cape Town. 2022. The Public Participation Process for the Cape Flats Aquifer. https://www.umvoto.com/our-work/public-participation-process/. Accessed Oct 2023.

[46] Décider Paris. 2023. Budget Participatif. https://decider.paris.fr/bp/jsp/site/Portal.jsp?page_id=10. Accessed August 2023.

[47] Décider Paris. 2023. Budget Participatif 2023 Resultats. https://www.paris.fr/pages/budget-participatif-2023-114-laureats-devoiles-25161. Accessed February 2024.

[48] Petitions. 2015. UK Government and Parliament. https://petition.parliament.uk/. Accessed 02/2024.

[49] S. Rafiee Rad and S. Roy. 2021. Deliberation and Single-Peakedness: A Computational Study. *Journal of Artificial Intelligence Research* 70 (2021), 1101–1130.

[50] John Rawls. 1958. Justice as Fairness. *The Philosophical Review* 67, 2 (1958), 164–194. http://www.jstor.org/stable/2182612

[51] W. David Ross. 1930. *The Right and the Good*. Oxford University Press, Oxford.

[52] Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin, United States.

[53] Kidjie Saguin. 2018. Why the poor do not benefit from community-driven development: Lessons from participatory budgeting. *World Development* 112

(2018), 220–232. https://doi.org/10.1016/j.worlddev.2018.08.009

[54] M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, R. Pinho Dos Santos, H. Vinagreiro Alves, E. Vecchione, and L. Scheunemann. 2021. *Ethics by Design and Ethics of Use Approaches for Artificial Intelligence.* Technical Report KJ-NA-30800-EN-N. Publications Office of the European Union. https://doi.org/10.2760/349527

[55] M Scharfbillig, L Smillie, D Mair, M Sienkiewicz, J Keimer, R Pinho Dos Santos, H Vinagreiro Alves, E Vecchione, and L Scheunemann. 2021. *Values and Identities - a policymaker's guide.* Scientific analysis or review KJ-NA-30800-EN-N (online),KJ-NA-30800-EN-C (print),KJ-NB-30800-EN-Q. European Comission, Luxembourg (Luxembourg). https://doi.org/10.2760/349527(online),10.2760/022780(print),10.2760/059689

[56] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 2307–0919.

[57] Amartya Sen. 2017. *Collective Choice and Social Welfare.* Harvard University Press, Cambridge, MA. https://doi.org/10.4159/9780674974616

[58] Marc Serramia, Maite Lopez-Sanchez, Stefano Moretti, and Juan A. Rodriguez-Aguilar. 2023. Building rankings encompassing multiple criteria to support qualitative decision-making. *Information Sciences* 631 (2023), 288–304. https://doi.org/10.1016/j.ins.2023.02.063

[59] Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Stefano Moretti. 2024. Value Alignment in Participatory Budgeting. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1692–1700.

[60] Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Manel Rodriguez, Michael Wooldridge, Javier Morales, and Carlos Ansotegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. IFAAMAS, Richland, SC, 1294–1302.

[61] Marc Serramia, Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. 2023. Encoding Ethics to Compute Value-Aligned Norms. *Minds and Machines* 33, 4 (2023), 761–790. https://doi.org/10.1007/s11023-023-09649-7

[62] Luciano C Siebert, Enrico Liscio, Pradeep K Murukannaiah, Lionel Kaptein, Shannon Spruit, Jeroen Van Den Hoven, and Catholijn Jonker. 2022. Estimating value preferences in a hybrid participatory system. In *HHAI2022: Augmenting Human Intellect*. IOS Press, Ansterdam, Netherlands, 114–127.

[63] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perello-Moragues. 2019. Value alignment: a formal approach. In *Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS*. IFAAMAS, Montreal, Canada, 15.

[64] John Smith. 2010. *Advanced Techniques in Forecasting.* Academic Press, New York, NY.

[65] Cass R. Sunstein. 2002. Deliberation and Polarization. *Political Philosophy* 3, 1 (2002), 1–24. https://doi.org/10.1002/9780470693711.ch1

[66] The Citizens' Assembly. 2017. The Eighth Amendment of the Constitution. https://2016-2018.citizensassembly.ie/en/The-Eighth-Amendment-of-the-Constitution/. Accessed: 2025-02-10.

[67] Steven R. Wilson, Yiting Shen, and Rada Mihalcea. 2018. Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In *Social Informatics*, Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov (Eds.). Springer International Publishing, Cham, 455–470.