A Computational Framework for Human Values

Nardine Osman
Artificial Intelligence Research Institute (IIIA-CSIC)
Barcelona, Catalonia, Spain
nardine@iiia.csic.es

Mark d'Inverno
Goldsmiths, University of London, UK
Artificial Intelligence Research Institute (IIIA-CSIC)
dinverno@gold.ac.uk

ABSTRACT

There is an increasing recognition of the need to engineer AI that respects and embodies human values. The value alignment problem, which identifies that need, has led to a growing body of research that investigates value learning, the aggregation of individual values into the values of groups, the alignment of norms with values, and the design of other computational mechanisms that reason over values in general. Yet despite these efforts, no foundational, computational model of human values has been proposed. In response, we propose a model for the computational representation of human values that builds upon a sustained body of research from social psychology.

KEYWORDS

human values, value representation, formal modelling

1 INTRODUCTION

Across governments, industry and the general public, there is an increasing recognition of the urgency for ethical approaches to AI (as evidenced by the numerous publications of ethics guidelines, e.g. [7, 8, 14, 15, 42]). In academia, a growing body of research investigates the role of human values in designing ethical AI [5, 13, 36, 44]. Indeed, one of our leading AI research luminaries, Stuart Russell, believes the overarching goal of AI should change from "intelligence" to "intelligence provably aligned with human values" [36]. This call to arms gave birth to the *value alignment problem.*

This challenge of engineering values into AI in response to the value alignment problem has resulted in a range of research areas: how AI learns human values [19–21, 45]; how individual values can be aggregated to the level of groups [18]; how arguments that explicitly reference values can be made [3]; how decision making can be value-driven [6, 10, 11]; how online institutions can ensure value-aligned behaviours in hybrid communities [27, 28]; and how norms are selected to maximise value-alignment [26, 39, 40].

Yet despite these efforts, no formal model of values exists that provides a concrete foundational platform from which data structures and algorithms can be designed to build AI architectures that address the value-alignment problem. In response, we present such

a model and set out the following guiding principles: (i) employ a formal language to be precise and a foundation for proof and algorithmic development (e.g. [12, 22]) (ii) ensure that our formal components lend themselves to data structure and algorithmic design (iii) subsume established concepts in AI research as much as possible, and (iv) make every effort to draw upon the wealth of work from within the humanities and social sciences (especially social psychology [34]). With this in hand, we have a better opportunity to address Russell's challenge of moving from "intelligence" to "intelligence aligned with human values" by supporting agreement on foundational and explicit models and properties of values.

Our model is presented in four subsections in Section 2. The first is a formal definition of values and value taxonomies. The second models the values of individual agents and groups of agents (both artificial and human), such as with many online communities. In the third, we extend the model to incorporate the changing values of individuals and communities over time. Finally, we address the problem of ensuring the extent (or degree) to which the behaviours of individuals or communities are aligned with an agreed set of values; the *value alignment problem.* Our ongoing work with medical doctors at Hospital del Mar, Barcelona, is making use of our proposal to interrogate and model their four overarching bio-ethical values [2] (beneficence, non-maleficence, autonomy, and justice) and so develop AI that provides feedback on the alignment of medical decisions leading to better value-aware decision-making in the hospital.

Each of those four subsections is divided into three parts: our proposal, a discussion of implementation choices, and a running example that embodies a concrete implementation choice. The running example should support the reader in understanding the development of our foundational proposal. We then close with a reflection on the contributions and further work (Section 3) along with some concluding remarks (Section 4).

## 2 A FORMAL MODEL FOR VALUE REPRESENTATION

Our model for value representation provides the foundations for new computational mechanisms that reason over values, enabling artificial systems to make value-aware decisions explicitly.

2.1 What are values? A computational approach

Our approach —aligned with Schwartz's Theory of Basic Values [37]— views values as human abstract concepts that guide behaviour and whose exact meaning and interpretation vary both with the current context and over time. However, a concrete computational representation of values requires a machine to reason with values. For example, while we might talk about fairness in general, for a specific application supporting a mutual aid community, fairness might be understood as "one does not
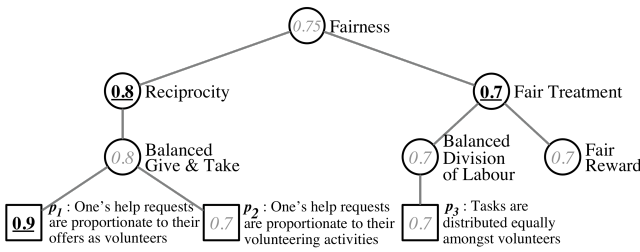
Figure 1: Abstractions for the value fairness: numbers indicate node importance, specified (black) and deduced (gray)

ask for help more than what they volunteer.". That is, the abstract concept of fairness acquires a concrete definition (grounding semantics) through a *property* whose satisfaction (or degree of satisfaction) can be automatically verified in a given system (more details are provided in Subsection 2.1.2). In this case, the property specifies that you cannot ask for more help than you volunteer.

Of course, there may be different levels of abstraction and grounding semantics for a value. For example, say the application was to be adopted by another community where volunteers explicitly support older people. The new community might find the old view of fairness —that one does not ask for help more than what they volunteer— unsuitable because this community expects to have volunteers that support older people, and they usually only ask for help with their day-to-day tasks without volunteering. As such, this new community should state that fairness is not about balanced give and take but about balanced workload division amongst the volunteers. Here, we encounter different levels of abstraction and grounding properties, illustrated in Figure 1.

The top node presents the abstract concept of fairness; the middle nodes present different concepts of fairness that are more specific than the top node, highlighting different abstraction levels and interpretations; and the bottom nodes present properties that ground the semantics of abstract concepts and allow for a computational evaluation of values. We use square nodes for nodes that ground the semantics of abstract concepts.

This interplay between the abstract and the grounding semantics of values is reflected through the relations between nodes, as shown in Figure 1. Different abstraction levels for value concepts may exist, along with different grounding semantics enabling computational approaches over values. Another key concept from Schwartz is *value priority*, which determines how *important* a value is for individuals or groups. In other words, it is not just the semantics of fairness that influences the behaviour of an individual or a group, but how important fairness as a value is for that individual or group. We model *value importance* by attaching a measure of *importance* to each node of Figure 1. We propose that values are computationally modelled as taxonomies, where a high-level value concept becomes more specific as one travels down any taxonomy path, becoming concrete and computational at leaf nodes. This is consistent with research in value-sensitive design [43] on value change taxonomies, as well as with Schwartz and a wide range of research from social psychology [34, 37].

Furthermore, using taxonomies allows for easy navigation between abstract and concrete grounding semantics of values, which

supports humans in understanding values and reflecting on and deliberating about them. It may be argued that all a machine requires are the property nodes to ensure behaviour alignment with values (as illustrated in Subsection 2.4). However, in the case of having a new community emerging, such as the community of volunteers supporting older people, we saw how a new community might have its local and distinct view of the fairness value. A taxonomy representation makes the deliberation process on the different semantics of fairness and the evolution of these semantics over time and context possible. We expect semantics to continuously evolve with the requirements of new interactions, as detailed in Subsection 2.2. In addition, we believe a taxonomy can support the deliberation process over values and their importance. Definition 1 presents our proposed taxonomy-based value representation.

DEFINITION 1 (VALUE TAXONOMY). *A value taxonomy $\mathcal{V} = (N, E, I)$ is defined as a directed acyclic graph, where:*

(1) *The set of nodes $N = N_l \cup N_\phi$ represents value concepts, and it is composed of two types of nodes: i) those that are specified through labels, with $N_l \subset L$ representing the set of label nodes and L is the set of all value labels representing abstract value concepts like 'fairness' or 'reciprocity'; and ii) those that are specified through concrete properties, with $N_\phi \subset \Phi$ representing the set of property nodes and Φ is the set of all value properties whose satisfaction can be automatically verified at different world states, such as having the number of times one offers help in a mutual aid community larger than the number of times one asks for help.*

(2) *The set of edges $E : N \times N$ is a set of directed edges $(n_p, n_c) \in E$ that represent the relation between value concepts n_p and n_c (the parent and child nodes, respectively) illustrating that the value concept n_p is a more general concept than n_c.*

(3) *The importance function $I : N \rightarrow COD_\perp$ either assigns an importance measure from the codomain COD to value concepts in N, or assigns $\perp$ to value concepts when their importance measure is undefined.*

Note that we specify the value taxonomy as a directed acyclic graph instead of the more traditional taxonomy tree because one value concept (node) may have more than one parent node. For example, the value 'equal treatment' in the taxonomy of Figure 2 can act as a more specific concept for both the 'social justice' and 'equality' values (where the social justice and equality values and their being examples of the universalism value are taken from Schwartz' value hierarchies [37]).
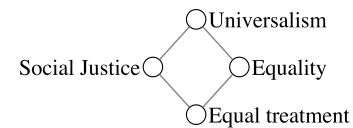
Figure 2: Value concepts with more than one parent node

We also require property nodes to be restricted to leaf nodes. In other words, the concrete grounding semantics of a value concept (specified as a property node) cannot be more general than another value concept (specified as another property or label node). For

example, in Figure 1, the property of having one's help requests proportionate to their volunteering offers cannot be more general than abstract concepts such as balanced give and take. This requirement is defined formally as follows: $\nexists\ (n_p, n_c) \in E \cdot n_p \in N_\phi$. We introduce this condition to simplify the construction and interpretation of value taxonomies. Future work may relax this requirement if property subsumption mechanisms are introduced to ensure parent nodes subsume children nodes.

There needs to be coherence regarding value importance within a taxonomy. Whilst humans are not always coherent, AI systems can be, and so can support humans to be more so. We express what we mean by coherence by requiring that the importance of a parent node be consistent with the combined importance of all of its child nodes. For example, if the importance of all children nodes is low, then the importance of the parent node cannot be high, and vice versa. We formally define the coherence of value importance accordingly:

DEFINITION 2 (COHERENCE OF VALUE IMPORTANCE). *Importance within a value taxonomy* $\mathcal{V} = (N, E, I)$ *is said to be coherent if and only if, for all nodes* $n \in N$ *with children nodes (there exists* $(n, n_c) \in E$*), the importance of the parent node is an aggregation of the importance of its children nodes:*

$$\forall n \in \{n_p \mid (n_p, n_c) \in E\}\ \cdot\ I(n) = \underset{n' \in X_n}{\mathbf{A}}\ I(n') \tag{1}$$

where $X_n = \{n_c \mid (n, n_c) \in E\}$ *is the set of all children nodes of* n*, and* $\mathbf{A} : COD^m \to COD$ *is an aggregation function that takes a set of size* $m \in \mathbb{N}^*$ *of importance measures in COD (specified as* COD^m*) and returns an aggregation of those measures, where the aggregation also falls in the same range COD.*

The importance measures of some of the taxonomy nodes might either be provided manually by humans or learned from other sources, such as past interactions. A coherence mechanism is then needed to ensure importance measures are coherent. Then, propagation mechanisms could be constructed to calculate the importance of nodes that have not been provided, building on existing propagation mechanisms [31, 32]. Any propagation mechanism must respect the coherence of value importance within the taxonomy.

But what about the aggregation function $\mathbf{A}$? We argue that symmetry, idempotence and monotonicity are desirable properties to be held by any such function. For instance, when calculating a parent node's importance, ordering its children nodes' importance measures should not affect the aggregation (symmetry). Also, if we assume all children nodes of some parent node to have the same importance i, then we believe the parent node should share that importance too (idempotence). It should neither be more important nor less important than i. Finally, increasing the importance measure of the children nodes should not decrease the parent node's importance (monotonicity). Formal definitions of these properties follow.

PROPERTY 1 (SYMMETRY OF AGGREGATION). *The aggregation function* $\mathbf{A}$ *is symmetric if, for all sets of importance measures* $\lambda \in COD^m$ *and all permutations* $\pi \in \Pi_\lambda$ *of those sets (where* Π_λ *is the set of all permutations of* λ*), we have:* $\mathbf{A}(\lambda) = \mathbf{A}(\pi)$

That is, the order of aggregated values does not matter.

PROPERTY 2 (IDEMPOTENCE OF AGGREGATION). *An aggregation function* $\mathbf{A}$ *is idempotent if, for all* $i \in COD$*, we have:* $\mathbf{A}(i, \ldots, i) = i$

That is, the aggregation of several instances of the same measure will return that same measure.

PROPERTY 3 (MONOTONICITY OF AGGREGATION). *An aggregation function* $\mathbf{A}$ *is monotonous if, for all sets of importance measures* $\lambda, \lambda' \in COD^m$*, we have:* $(\forall\ 0 < p \le m\ \cdot\ \lambda_p \le \lambda'_p)\ \Rightarrow\ \mathbf{A}(\lambda) \le \mathbf{A}(\lambda')$ *, where* λ_p *represents the element in position* p *of the set* λ*.*

That is, if the measures in λ are smaller than or equal to those in λ' (per position), then the aggregation of the measures in λ should be smaller than or equal to the aggregation of those in λ'.

Our requirements for Properties 1–3 help define the type of aggregation operator. First, from [24, p. 14], we know that the idempotence and monotonicity properties imply compensativeness. Compensative operators are aggregation operators that are neither conjunctive nor disjunctive. They are limited between the min and max, which are the bounds of the t-norm and t-conorm families. This implies that for any set of importance measures (for all $\lambda \in COD^m$), the aggregation will fall between the minimum and maximum measures in that set: $\min \lambda \le \mathbf{A}(\lambda) \le \max \lambda$.

We believe falling between the minimum and maximum is appropriate for our setting: a parent node should not be more important than its most important child node nor less important than its least important child node. Furthermore, from [24, p. 13], we also know that compensative aggregation operators that satisfy the symmetry and monotonicity properties are averaging operators. As such, we propose $\mathbf{A}$ to be an averaging operator, though the exact choice of this operator is left for implementation.

2.1.1 Implementation Choices.

Specifying property nodes. We use properties to specify the grounding semantics of values, as properties have traditionally been used to describe the world state, and their satisfaction can be computed [41]. The exact language used for specifying properties is an implementation decision. Our proposal is agnostic regarding any choice of implementation language or algorithmic design. It can encompass any theory written in propositional, first-order, deontic or modal logic.

We note that the use of properties may initially appear to embody a consequentialist view, where only the outcome of behaviour is what matters. However, in our model, properties can be designed to evaluate not only action outcomes but also attached to actions themselves (such as whether an action is permitted, e.g. lying is never acceptable). The expressiveness of the chosen language for implementing properties plays a significant role here. In this paper, we limit the examples to propositional logic and, more specifically, simple propositions to improve readability.

Choosing the codomain of value importance. Concerning the importance measure of a value concept, the choice of the codomain COD that evaluates this importance I is also an implementation decision. Schwartz, for instance, used the range $\{-1, 0, 1, 2, 3, 4, 5, 6, 7\}$ for people to specify the importance of a value, such as 'equality' (equal opportunity for all) or 'pleasure' (gratification of desires), as a guiding principle in their lives [37], where -1 represents opposing a value, 0 represents considering the value to be non-important,

and positive numbers represent different degrees of supporting a value —'supreme importance', 'very important', 'important', and so on.

Alternative approaches, such as defining importance as a partial or total order, may also be considered. Schwartz argues that this order (the relevant importance [37]) should be used when reasoning about values, where the order is deduced from the importance assigned by people. A partial order may sometimes be more intuitive for humans to provide, as opposed to giving numerical importance measures. When this is the case, human stakeholders may use partial orders to specify importance measures and conversion mechanisms may be used to translate partial orders into numeric values. In general, the implementation choice must depend on the specific application's requirements. In this paper's example, we set the codomain to $[-1, 1]$ as reasoning with numbers is computationally easier than reasoning with partial orders. Furthermore, the range $[-1, 0]$ is more intuitive for describing opposition to a value. This expands Schwartz's opposition measures, as there may also be degrees when opposing a value the same way there are degrees in supporting a value.

Ensuring the coherence of value importance. As for the aggregation function $\mathbf{A}$ that ensures the coherence of value importance within a taxonomy, different averaging operators may be investigated. In this paper, we propose a simple average:

$$I(n_p) = (\sum_{n_c \in X_{n_p}} I(n_c)) / (|X_{n_p}|) \tag{2}$$

Recall that we assume the importance measure of some nodes to either be provided manually or learnt. A propagation mechanism may then be implemented to calculate the importance of nodes that have not been manually provided or learnt, all while ensuring the overall coherence of value importance within the taxonomy (Definition 2). An example propagation mechanism can be found in [29, Alg. 1].

2.1.2 The Running uHelp Example.
We'll consider the value *fairness* for our running example. This value has emerged as an important value in the participatory design meetings with potential users of the uHelp app.[1] The uHelp app is a social networking app that allows people to find help within their social network with their day-to-day activities, such as finding someone to substitute for them at work tomorrow or someone to lend them some chairs for a party [16, 17, 30]. We choose not to confine our model to predefined value systems, like Schwartz's infamous set of universal human values [37], but opt for values that may emerge from different stakeholders.

Recognising relevant high-level values (root nodes) may be achieved either through a bottom-up approach, in which relevant values emerge from discussing the context with the relevant stakeholders, or through a top-down approach, in which a given stakeholder specifies the relevant values from prior knowledge. The choice is usually use case dependent. For example, in other interactions with doctors and firefighters, relevant values were predefined

---

[1]The uHelp app (https://uhelpapp.com) was developed at IIIA-CSIC in collaboration with the first author of this article. It is available on Google Play (https://play.google.com/store/apps/details?id=es.csic.iiia.uhelpapp) and Apple Store (https://itunes.apple.com/es/app/uhelp/id1089461370).

by the relevant institutions and provided in a top-down approach. Once this has taken place the rest of the taxonomy emerged from contemplating those values and how they are understood in a given context. As new contexts emerged over time, the taxonomy was updated to consider the requirements of these new contexts. For example, fairness was initially understood as balanced give and take for a mutual aid community of single mothers. In other words, the users wanted to avoid greedy members asking for help and never offering it. But when, later on, we discussed uHelp with a community providing support to older people, it became clear that a balanced give-and-take was against their values. A new understanding of fairness emerged: the balanced division of labour. In other words, while some community members will ask for help and others will provide it, the workload among those giving help should be balanced. Figure 1 illustrates the evolving value taxonomy for fairness from uHelp's perspective. Next, we define some of the grounding semantics of those abstract concepts, i.e. defining the property leaf nodes. Recall that property nodes are presented as squares, whereas label nodes are shown as circles.

For illustrative purposes, we provide two different computational approaches defining balanced give and take using properties p_1 and p_2. The first states that one's help requests are proportionate to the number of times the user offered help, whereas the second states that one's help requests are proportionate to the number of times the user was chosen to help (because not all those offering help get selected). One straightforward approach to specifying p_1 and p_2 is through the ratio of requests to offers/volunteering being greater than 1, as illustrated in property definitions 3 and 4. Of course, we provide simple properties for better readability, but we note that properties can get as complex as the language allows (remember that the choice of language for specifying properties is an implementation choice, and one may choose to work with highly expressive languages).

$$p_1 \stackrel{def}{=} (\#_{requests}) / (\#_{offers}) > 1 \tag{3}$$

$$p_2 \stackrel{def}{=} (\#_{requests}) / (\#_{volunteering}) > 1 \tag{4}$$

The two different properties for balanced give and take illustrate how different semantics may be provided for the same abstract value. In this specific case, one can imagine p_2 to have been initially defined, but after some interactions, some users are never chosen despite their volunteering, and as such, the system prevents them from asking for help so that balanced give and take is respected. To compensate for this, p_1 gets added to the taxonomy so that balanced give and take considers the offers for help instead of being chosen. This illustrates how the taxonomy may evolve by learning from experience what the best grounding semantics (property) of a given abstract concept are.

The computational approach, or the property node, describes the balanced division of labour through property p_3, which states that tasks are equally distributed amongst volunteers. One approach to specifying p_3 is through the difference between the uniform distribution U and the distribution D of the numbers of tasks assigned to each volunteer, where this difference should be smaller than a predefined threshold ϵ (property definition 5).

$$p_3 \stackrel{def}{=} difference(D, U) < \epsilon \tag{5}$$

where the difference between two distributions may be calculated using approaches like the Kullback–Leibler divergence [23] or the earth mover's distance [35].

This taxonomy does not specify the property node providing the grounding semantics for fair reward. Indeed, some abstract concepts may remain abstract for some time. Furthermore, we note that while stakeholders discuss and agree on abstract value concepts, engineers must define the properties that ground the semantics of those values. Alternatively, AI may support constructing and updating these value taxonomies based on emerging value learning mechanisms [4, 20].

A crucial feature of the value taxonomy is the importance of nodes. We imagine each context to have its own taxonomy with its own importance measures (see Subsection 2.2). As such, an adapted uHelp taxonomy for fairness will be made available for the community of single mothers and another for the community supporting older people. However, whether importance measures are available for the general taxonomy of uHelp is application-dependent.

Getting stakeholders to specify the importance of all nodes is usually challenging, even if importance is provided through a partial order. As such, we give an example where importance is assigned to a subset of these nodes (Figure 1), where we represent importance measures explicitly provided through bold underlined numbers. From those, our propagation mechanism (see [29, Alg. 1]) then calculates possible coherent importance measures for the remaining nodes (the non-underlined numbers). We have provided the value taxonomy of one uHelp value: fairness. We imagine applications to have several relevant values and, thus, several such taxonomies.

2.2 How do values change with context? Context-based value taxonomies

Our stance is that values are context-dependent. This is the stance of value-sensitive design [43], as well as the stance of many social psychologists before them [34]. We argue that we all have a general view of what a value is, defined through its value taxonomy, and this view evolves with our experiences, where new nodes (label-based and property-based) are continuously added. If a new context requires adding new nodes, those are added to the general taxonomy. This is how, in general, taxonomies evolve with experiences.

Suppose a new context necessitates eliminating existing nodes. This is achieved by setting their importance to zero for that specific context (if zero is the neutral-based measure of the chosen codomain). Moving from a general taxonomy to a context-based one is implemented by simply revisiting the importance measures of the general taxonomy (whether they were defined or not), making some nodes or branches more prominent than others. We note that if property-based leaf nodes did not exist for a prominent branch, they must be added to the general taxonomy. Otherwise, a computational approach considering those branches will not be possible. Our definition of a context-based value taxonomy is presented next.

DEFINITION 3 (CONTEXT-BASED VALUE TAXONOMY). *A context-based taxonomy $\mathcal{V}_c = \{N, E, I_c\}$ is an alteration of a general taxonomy $\mathcal{V} = \{N, E, I\}$ where the importance of nodes are updated for the given context. The importance of nodes in the context-based taxonomy $\mathcal{V}_c$ is independent of the importance of those nodes in the general*

taxonomy $\mathcal{V}$. The function calculating the importance of nodes with respect to a context c is defined as $I_c : N \rightarrow COD_\perp$, where $COD_\perp$ defines the union of COD with the undefined variable ($\perp$).

2.2.1 Implementation Choices.

Constructing Context-Based Taxonomies. Different mechanisms for evaluating the importance of nodes in context-based taxonomies may be developed. One possible implementation (see [29, Alg. 2]) would take a bottom-up approach where only the importance of property nodes for the given context is evaluated (regardless of how they are obtained, whether they are provided manually by stakeholders or learnt from similar past experiences of similar contexts). This approach aims to better assess the importance of concrete property nodes in specific contexts. Then, the importance of the remaining nodes in the taxonomy is calculated by propagating the importance measures of the property nodes across the taxonomy, as described in Subsection 2.1.1.

Alternative implementations may be investigated. Rather than starting with property nodes and following a bottom-up approach, a top-down approach may be implemented to assess abstract concepts regardless of their grounding semantics. Other domains suggest a mix. Finally, we note that inconsistencies may arise between general and context-based taxonomies. This is expected and entirely normal. For example, while reciprocity might be considered very important as an abstract concept for the uHelp app, it might be regarded as less important for a specific context, such as for the community of volunteers supporting older people (see Subsection 2.2.2).

Visualising Taxonomies. In addition to constructing context-based taxonomies through updating importance measures and for improved visualisation, we propose removing parts of the taxonomy deemed irrelevant for a given context. This step is optional but may be useful when visualising growing taxonomies for a given context. This also helps stakeholders quickly spot the relevant nodes for this context.

Different approaches may be followed when deciding which branches are relevant hiding irrelevant details. For example, one approach (see [29, Alg. 2]) is to maintain the branches that lead to relevant property nodes and eliminate those that lead to irrelevant property nodes, where the importance of nodes dictates. Relevance may be domain dependent and so determined by the chosen codomain. In this implementation of our model, irrelevant nodes are those with zero importance.

2.2.2 The Running uHelp Example.
While Figure 1 has presented uHelp's general taxonomy for fairness, we present examples of context-based taxonomies for different uHelp communities in this subsection. As illustrated in Subsection 2.1.2, the first community to adopt uHelp was the community of single mothers (whose context is specified as c_s). For them, fairness implied a balanced give and take. We could imagine that at the beginning, both the general taxonomy and the context-based taxonomy for the c_s community were defined through the taxonomy of Figure 3a. Then, as another context emerges, which is that of the community supporting older people (whose context is specified as c_e), new nodes were added to the general taxonomy, resulting in the Figure 1 taxonomy.

Armed with this new general taxonomy for fairness, the community supporting older people decides that the requirement of
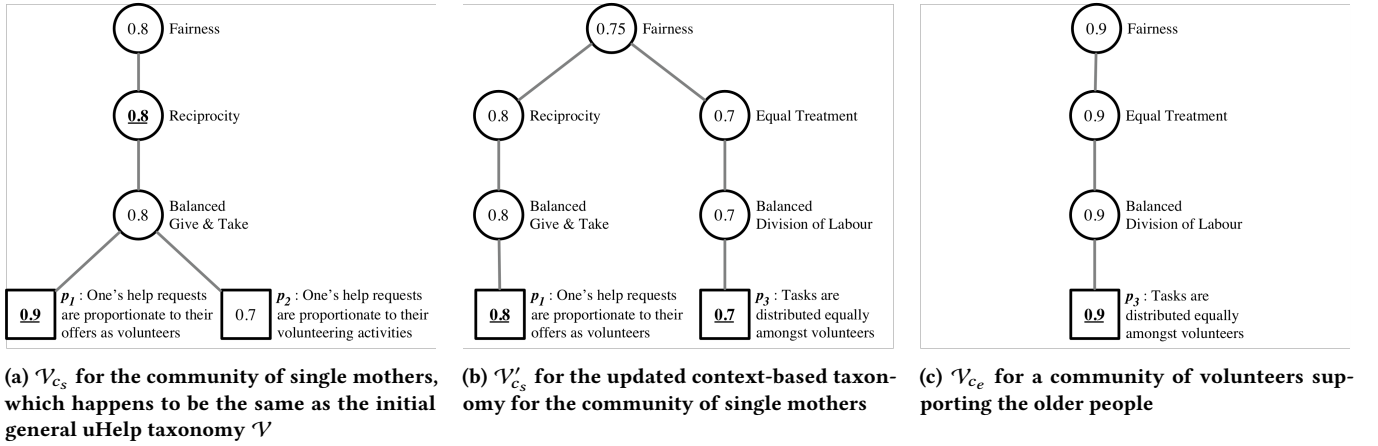
(a) $\mathcal{V}_{c_s}$ for the community of single mothers, which happens to be the same as the initial general uHelp taxonomy $\mathcal{V}$

(b) $\mathcal{V}'_{c_s}$ for the updated context-based taxonomy for the community of single mothers

(c) $\mathcal{V}_{c_e}$ for a community of volunteers supporting the older people

Figure 3: Different context-based value taxonomies for fairness in uHelp

having a proportionate number of help requests with respect to the number of help offers (p_1) or volunteering (p_2) is irrelevant. In contrast, the requirement that labour should be divided in a balanced way amongst all volunteers is of utmost importance. So the importance of the property nodes of Figure 1 are specified as follows:

$$I_{c_e}(p_1) = 0 \quad ; \quad I_{c_e}(p_2) = 0 \quad ; \quad I_{c_e}(p_3) = 0.9$$

The context-based taxonomy of this new community is visualised (in our approach) by eliminating all branches leading to property nodes with zero importance, leaving us with one single branch leading to property node p_3. The resulting taxonomy is presented in Figure 3c. The importance of each of the upper nodes inherits the importance of the property node, following our proposed propagation mechanism (see [29, Alg. 1]). With this new general taxonomy (Figure 1), the single mother's community decides to revise their context-based taxonomy. They like the new grounding semantics that have emerged and determine that it is important that requests should be both proportionate to offers (p_1), and equally distributed amongst volunteers (p_3). The new importance then reflects this new view measures they assign to the property nodes of Figure 1:

$$I_c(p_1) = 0.8 \quad ; \quad I_c(p_2) = 0 \quad ; \quad I_c(p_3) = 0.7$$

Their new taxonomy is updated and visualised by eliminating branches that lead to property nodes with zero importance. This results in Figure 3b. As before, the importance of the upper nodes is inherited from the importance of the property nodes.

2.3 Who holds values? Individuals vs collectives

Entities hold values: values do not exist on their own. In other words, there is no universal value taxonomy for fairness. Different people or groups will hold different views on values, leading to different taxonomies. We use the notation $\mathcal{V}^x$ to represent the value taxonomy held by entity x, where x may represent an individual i_j or a collective $\{i_1, \ldots, i_n\}$. We use the word collective to describe a group of interacting individuals, which may be a community, an organisation, an institute, a society, a culture, etc. When a collective holds a value taxonomy, this is understood as the value taxonomy

describing the values of the collective as a whole and not its individuals (or interacting members). Individuals may not all have their value taxonomy aligned with the collective's. The issue of how the collective agrees on its taxonomy is left for future work. To simplify notation, in the rest of the paper, we drop the x from $\mathcal{V}^x$ when it is clear who holds the taxonomy. As autonomous agents, humans (and future software agents) do not only understand their own values or the values of the collectives they belong to, but they also observe others. They can form beliefs about the values of others. We use the notation $\mathcal{V}^{x>y}$ to represent what x believes are the values of y, where both x and y may represent an individual or a collective.

2.3.1 Implementation Choices. The issue here is implementing mechanisms that take the value taxonomies of individuals and compute a taxonomy of these individuals as a collective. This is a complex task, and the domain dictates how collective values are specified. For example, the company developing an assistive robot may pre-define the value taxonomy governing the behaviour of that robot. In a hospital, an elected board of medics convene to agree on the hospital's values collectively. Yet. in other cases, such as the uHelp application, one can imagine the community of users coming together to vote on their values. In other words, the rules dictating whose view should be considered when specifying the value taxonomies of a collective are domain dependent.

Different mechanisms may be developed to construct collective values from individual ones. For example, negotiation and argumentation mechanisms can help individuals reach collective agreements on their values. Computational social choice may also be used to aggregate individual values into collective ones. We point the interested reader to ongoing work on this topic [18], where the aggregation takes into consideration various ethical principles, such as utilitarian (maximum utility) or egalitarian (maximum fairness).

2.3.2 The Running uHelp Example. This subsection highlights that both individuals and collectives may hold value taxonomies. For example, imagine three uHelp members from the community of single mothers who hold the taxonomies of Figures 4a–4c. Of course, more members would exist in reality, each with their own taxonomy, but for simplicity, we show the taxonomies of three individuals
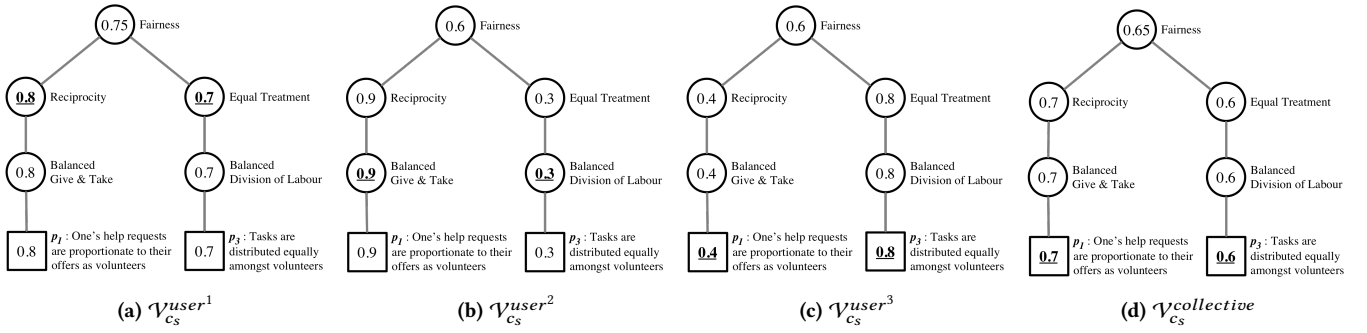
Figure 4: Individual and collective value taxonomies for fairness in uHelp's community of single mothers

only. Given these taxonomies, the social choice mechanism of [18] can then be used to compute the importance of each property node for the collective. With those, a propagation mechanism (such as that of [29, Alg. 1]) can then be used to propagate those to the rest of the tree. The resulting taxonomy is shown in Figure 4d.

2.4 Why hold values? Value-alignment problem

Values are one of the main motivators of behaviour [34, 37]. Assessing the alignment of behaviour with values has been the main objective of the work on values in AI to ensure value-aligned behaviour. The property-based nodes of a value taxonomy introduce the foundations for linking abstract value concepts to concrete computational constructs that can help formally assess the alignment of behaviour with those values. The value alignment of an entity's behaviour becomes the degree of satisfaction of the property-based value nodes of the relevant taxonomy that this behaviour brings about. In Definition 3 we have seen how importance is assigned to different taxonomy nodes.[2] The evaluation of value alignment should take into consideration these importance measures: the more important a property-based node is, the higher its satisfaction contributes to the value alignment of the behaviour being evaluated, and vice versa. The alignment of an entity e's behaviour with a context-based value taxonomy $\mathcal{V}_c$ is then defined accordingly:[3]

$$\mathcal{A}(e, \mathcal{V}_c) = \bigoplus_{p \in N_{\phi,c}} f(sd(p, e), I_c(p)) \qquad (6)$$

where $N_{\phi,c}$ represents the property nodes of the taxonomy $\mathcal{V}_c$, $sd(p, e)$ represents the degree of satisfaction of property p with respect to the behaviour of entity e, and $I_c(p)$ represents the importance of the property-node p within the context-based value taxonomy $\mathcal{V}_c$. The function f is used to factor in the importance of property nodes when considering their degree of satisfaction, whereas $\bigoplus$ is used to aggregate the degree of satisfaction of all property nodes in $\mathcal{V}_c$ (with value importance factored in).

2.4.1 *Implementation Choices.* One question is calculating the satisfaction of property nodes $sd(p, e)$. In other words, given an entity e, how do we assess to what degree the behaviour of e results in the satisfaction of property p? This requires knowledge about how e behaves, and different implementation approaches for specifying this knowledge can be followed. For example, suppose e is a complex system of communicating entities. In that case, e's model (usually specified via a process calculi) will describe its behaviour through a labelled transition system where the satisfaction of specific properties at different states [41] can be evaluated. If e is a normative system, then the norms can help map out the state diagram of the possible interaction outcome and evaluate the satisfaction of relevant properties accordingly [1, 9, 25]. If e was an agent with a BDI model, then BDI reasoning mechanisms can help assess the degree to which specific properties will be satisfied by e's behaviour [33]. In summary, a model of e describing its behaviour is necessary to assess $sd(p, e)$. This issue has already been addressed in many fields, as illustrated above. To ensure our proposal is not limited to one modelling choice, we omit the choice of modelling e's behaviour and assume the degree of satisfaction $sd(p, e)$ to be attainable.

Returning to the alignment function $\mathcal{A}$, there are other implementation choices, such as the choice of the function f that factors in the importance of a property node and the aggregation operator $\bigoplus$. In this paper, we propose a straightforward implementation that follows a weighted average approach:

$$\mathcal{A}(e, \mathcal{V}_c) = \left(\sum_{p \in N_{\phi,c}} I_c(p) \cdot sd(e, p) \right) / (|N_{\phi,c}|) \qquad (7)$$

Assuming the range of value importance I to be $[-1, 1]$, and degree of satisfaction $sd(e, p)$ a number with the range $[0, 1]$, then the range of $\mathcal{A}$ becomes $[-1, 1]$ where negative results describe the degree of misalignment (or an alignment with detested values) and positive results describe the degree of alignment with aspired values.

2.4.2 *The Running uHelp Example.* Let us consider the context-based value taxonomy $\mathcal{V}'_{c_s}$ of Figure 3b for a mutual aid community. The concrete definitions of p_1 and p_3 (property definitions 3 and 5) illustrate what it means, computationally, for the behaviour of some entity to be aligned with the value 'fairness' in this context. Next, we illustrate how the exact degree of satisfaction of p_1 and p_3 can be computed according to these definitions. Equation 8 formally states that the degree of satisfaction of p_1 is the actual ratio of

---

[2]Value-alignment will be assessed within different contexts, which is discussed in Subsection 2.2 to context-based value taxonomies.

[3]As before, context-based value taxonomies are expected to describe the values held by some entity. However, for the sake of simplifying notation, we drop the holder x (and possibly x's view of y's values, if that was the case) and replace $\mathcal{V}^x_c$ with $\mathcal{V}_c$ (or $\mathcal{V}^{x>y}_c$ with $\mathcal{V}_c$). We also note that x (or even y) does not necessarily have to be the same entity e whose behaviour is being assessed. In other words, if $x = e'$, then this describes assessing how much e is aligned with the values of e'.

requests to offers, normalised to fall into the range $[-1, 1]$.

$$sd(e, p_1) = \begin{cases} (R - 1) \; / \; ((\max R) - 1) & \text{, if } R > 1 \\ R - 1 & \text{, otherwise} \end{cases} \quad (8)$$

where $R = \#requests/\#offers$ represents the ratio of requests to offers, and $\max R$ is the maximum possible value for R. While the range of R is $[0, \infty)$, a maximum value must be selected for our equations. We argue that $\max R$ is domain dependent and should be selected for each scenario. Equation 8 states that the degree of satisfaction is computed by normalising the ratio R of requests to offers to the range $[-1, 1]$. The degree of satisfaction of p_1 depends on how far is the ratio R from 1. The larger it is with respect to 1, the higher the degree of satisfaction. The closer it is to 0, the greater the degree of dissatisfaction.

Next, Equation 9 defines the satisfaction of property p_3 similarly by formally stating that the degree of satisfaction of p_3 is the actual difference between the uniform distribution U and the distribution of tasks over volunteers D, normalised to the range $[-1,1]$.

$$sd(e, p_3) = \begin{cases} 1 - (\Delta \; / \; \epsilon) & \text{, if } \Delta < \epsilon \\ (-(\Delta - \epsilon)) \; / \; ((\max \Delta) - \epsilon) & \text{, otherwise} \end{cases} \quad (9)$$

where $\Delta = difference\,(D, U)$ represents the difference between the distribution of tasks over volunteers (D) and the uniform distribution (U), and $\max \Delta$ is the maximum possible value for Δ. The range of Δ is $[0, \infty)$, but a maximum value must be selected for our equations. Again, we argue that $\max \Delta$ is domain dependent, and must be chosen for each scenario. Equation 9 states that the degree of satisfaction of p_3 depends on how far is the difference Δ from ϵ. The larger it is with respect to ϵ, the higher the degree of dissatisfaction. The closer it is to 0, the greater the degree of satisfaction.

Now, say a mutual aid community e provides incentives that motivate people to volunteer, and has norms that ensure tasks are spread as equally as possible over volunteers. Say the regimented norms result in a high degree of satisfaction for p_3, whereas the incentives result in a mediocre degree of satisfaction for p_1:

$$sd(e, p_1) = 0.5 \quad ; \quad sd(e, p_3) = 0.9$$

And say the importance of p_1 is set to be twice that of p_2:

$$I_c(p_1) = 1 \quad ; \quad I_c(p_3) = 0.5$$

Following Equation 7, the alignment of the mutual aid community e with its understanding of fairness $\mathcal{V}'_{c_s}$ (Figure 3b) becomes:

$$\mathcal{A}(e, \mathcal{V}_c) = 0.475$$

3 CONTRIBUTIONS AND FURTHER WORK

The novel contribution is a formal and computational model of human values that lays the foundations for algorithmic reasoning over them, which is strongly aligned with existing research from social psychology. Schwartz and Bilsky's five features that are recurrently mentioned in the literature to define values state that values "(1) are concepts or beliefs, (2) pertain to desirable end states or behaviours, (3) transcend specific situations, (4) guide selection or evaluation of behaviour and events, and (5) are ordered by relative importance" [38].

These features are shared by many social psychologists and social scientists [34], aligning with our value taxonomy proposal as follows. Values are abstract concepts (feature 1), specified through 'labels' like fairness, equality, etc. They are defined through desirable end states (feature 2), implemented via property nodes. The whole work on values is to guide behaviour (feature 4): our value-alignment mechanism assesses to what extent behaviour is aligned with selected values. Value importance is integral to our approach (feature 5), where node importance is critical for computing value alignment. While feature 3 states that values transcend specific situations, we argue that although value taxonomies do not change frequently, they do evolve. Here, we are more aligned with the work in value-sensitive design [43].

One major challenge is designing and constructing value taxonomies. The human values we want to specify and embed in the decision-making processes need to come from diverse human stakeholders, including users, designers, owners, and others directly or indirectly affected. One approach involves stakeholders explicitly detailing their values in a way that can be directly mapped to a formal model such as ours; another involves AI learning stakeholder values from their interactions. The first approach requires significant effort from humans, while the second is prone to errors in the learning mechanism. While impressive value learning [20] and value aggregation [18] mechanisms are being proposed, they are not error-free, and they do not deal with the complexity of value taxonomies. Introducing these taxonomies introduces new challenges, such as learning the importance of values, the relations between value nodes, the property nodes for some value concepts, and designing mechanisms to aggregate value taxonomies.

The second issue concerns the modelling process, necessitating representation choices that may bring limitations. We have introduced guiding principles that support our modelling decisions and aimed to consider theoretical notions that clarify implementation choices. Whilst we deliberately leave the choice of representation and implementation open for system and research development, we have made implementation choices in our running example.

4 CONCLUSIONS

We have contributed to the urgent challenge of building value-aligned AI by proposing a conceptually intuitive foundational model for human values. It allows for future computational reasoning and opens up opportunities to evidence how AI systems are *provably* aligned with human values. The approach is grounded in social psychology, subsumes existing AI research concepts, and is formal, making it a coherent and intuitive starting point for future interdisciplinary research investigation. To our knowledge, this is the first proposal for the formal representation of human values and moves beyond the state-of-the-art —which to date has defined values through labels [20, 39] or goals [26, 40]— by explicitly introducing notions of value importance, semantics, and relations.

REFERENCES

[1] Thomas Ågotnes, Wiebe Van Der Hoek, Juan A Rodríguez-Aguilar, Carles Sierra, and Michael J Wooldridge. 2007. On the Logic of Normative Systems. In *IJCAI*, Vol. 7. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1175–1180.

[2] Tom L. Beauchamp. 1979. *Principles of Biomedical Ethics*. Oxford University Press.

[3] Trevor Bench-Capon and Katie Atkinson. 2009. Abstract Argumentation and Values. In *Argumentation in Artificial Intelligence*, Guillermo Simari and Iyad Rahwan (Eds.). Springer US, Boston, MA, 45–64.

[4] Emanuele Brugnolia, Pietro Gravinoa, and Giulio Prevedellob. 2023. Moral values in social media for disinformation and hate speech analysis. Pre-Proceedings of the Value Engineering in AI Workshop (VALE) ECAI 2023. https://vale2023.iiia.csic.es/accepted-papers

[5] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. 2021. Trustworthy AI. In *Reflections on Artificial Intelligence for Humanity*. LNCS, Vol. 12600. Springer, 13–39.

[6] Kinzang Chhogyal, Abhaya C. Nayak, Aditya Ghose, and Hoa Khanh Dam. 2019. A Value-based Trust Assessment Model for Multi-agent Systems. In *Proc. of the 28th Int. Joint Conf. on Artificial Intelligence Organization*. IJCAI, 194–200.

[7] European Commission. 2019. Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[8] European Commission. 2021. Proposal for a Regulation of the European Parliament and of the council: laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM/2021/206 final, with Procedure Number 2021/0106/COD. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206.

[9] Stephen Cranefield and Michael Winikoff. 2011. Verifying social expectations by model checking truncated paths. *Journal of Logic and Computation* 21, 6 (2011), 1217–1256.

[10] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Proc. of the 26th International Joint Conference on Artificial Intelligence*. IJCAI, 178–184.

[11] Gennaro di Tosto and Frank Dignum. 2012. Simulating Social Behaviour Implementing Agents Endowed with Values and Drives. In *Multi-Agent-Based Simulation XIII - International Workshop, MABS, Revised Selected Papers (LNCS, Vol. 7838)*, Francesca Giardini and Frédéric Amblard (Eds.). Springer, 1–12.

[12] Mark d'Inverno and Michael Luck. 1997. Development and application of a formal agent framework. In *First IEEE Int. Conf. on Formal Engineering Methods*. IEEE Computer Society, 222–231. https://doi.org/10.1109/ICFEM.1997.630429

[13] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds Mach.* 30, 3 (sep 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[14] IEEE Standards Association. [n.d.]. The IEEE Global Initiative. https://standards.ieee.org/industry-connections/ec/autonomous-systems/.

[15] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[16] Andrew Koster, Jordi Madrenas-Ciurana, Nardine Osman, W. Marco Schorlemmer, Jordi Sabater-Mir, Carles Sierra, Dave de Jonge, Angela Fabregues, Josep Puyol-Gruart, and Pere Garcia-Calvés. 2012. u-Help: Supporting Helpful Communities with Information Technology. In *Proceedings of the First International Conference on Agreement Technologies, AT 2012*, Sascha Ossowski, Francesca Toni, and George A. Vouros (Eds.), Vol. 918. CEUR-WS.org, 378–392. https://ceur-ws.org/Vol-918/111110378.pdf

[17] Andrew Koster, Jordi Madrenas-Ciurana, Nardine Osman, W. Marco Schorlemmer, Jordi Sabater-Mir, Carles Sierra, Angela Fabregues, Dave de Jonge, Josep Puyol-Gruart, and Pere Garcia-Calvés. 2013. u-Help: supporting helpful communities with information technology. In *Proc of the Twelfth International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '13*, Maria L. Gini, Onn Shehory, Takayuki Ito, and Catholijn M. Jonker (Eds.). IFAAMAS, 1109–1110.

[18] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite López-Sánchez, and Juan A. Rodríguez-Aguilar. 2022. Towards Pluralistic Value Alignment: Aggregating Value Systems Through l_p-Regression. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 780–788.

[19] Ying Lin, Joe Hoover, Morteza Dehghani, Marlon Mooijman, and Heng Ji. 2018. Acquiring Background Knowledge to Improve Moral Value Prediction. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2018), 552–559.

[20] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. IFAAMAS, Richland, SC, 799–808.

[21] Hui Liu, Yinghui Huang, Zichao Wang, Kai Liu, Xiangen Hu, and Weijun Wang. 2019. Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System. *Applied Sciences* (2019).

[22] Michael Luck and Mark d'Inverno. 1995. Structuring a Z specification to provide a formal framework for autonomous agent systems. In *ZUM '95: The Z Formal Specification Notation*, Jonathan P. Bowen and Michael G. Hinchey (Eds.). Springer, Berlin, Heidelberg, 46–62.

[23] David J.C. MacKay. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

[24] J.-L. Marichal. 1998. *Aggregation Operators for Multicriteria Decision Aid*. Ph.D. Dissertation. Institute of Mathematics, University of Liège, Liège, Belgium.

[25] Nieves Montes, Nardine Osman, and Carles Sierra. 2022. A computational model of Ostrom's Institutional Analysis and Development framework. *Artificial Intelligence* 311 (2022), 103756. https://doi.org/10.1016/J.ARTINT.2022.103756

[26] Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems*, Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé (Eds.). ACM, 907–915.

[27] Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. 2022. Design Heuristics for Ethical Online Institutions. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV*, Nirav Ajmeri, Andreasa Morris Martin, and Bastin Tony Roy Savarimuthu (Eds.). Springer International Publishing, Cham, 213–230.

[28] Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. 2023. Addressing the Value Alignment Problem Through Online Institutions. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI*, Nicoletta Fornara, Jithin Cheriyan, and Asimina Mertzani (Eds.). Springer Nature Switzerland, Cham, 77–94.

[29] Nardine Osman and Mark d'Inverno. 2023. A computational framework of human values for ethical AI. *CoRR* abs/2305.02748 (2023). https://doi.org/10.48550/ARXIV.2305.02748 arXiv:2305.02748

[30] Nardine Osman, Bruno Rosell, Carles Sierra, Marco Schorlemmer, Jordi Sabater-Mir, and Lissette Lemus. 2023. uHelp: intelligent volunteer search for mutual help communities. *CoRR* abs/2301.11112 (2023). https://doi.org/10.48550/arXiv.2301.11112 arXiv:2301.11112

[31] Nardine Osman, Carles Sierra, and Jordi Sabater-Mir. 2010. Propagation of Opinions in Structural Graphs. In *Proc. of the 19th European Conference on Artificial Intelligence, ECAI 2010 (Frontiers in Artificial Intelligence and Applications, Vol. 215)*, Helder Coelho, Rudi Studer, and Michael J. Wooldridge (Eds.). IOS Press, 595–600.

[32] Nardine Osman, Carles Sierra, Jordi Sabater-Mir, Joseph R. Wakeling, Judith Simon, Gloria Origgi, and Roberto Casati. 2010. LiquidPublications and its technical and legal challenges. In *Intelligent Multimedia: Managing Creative Works in a Digital World*. Legal Information and Communication Technologies, Vol. 8. European Press Academic Publishing, Florence.

[33] Anand S Rao and Michael P Georgeff. 1998. Decision procedures for BDI logics. *Journal of logic and computation* 8, 3 (1998), 293–343.

[34] Meg J. Rohan. 2000. A Rose by Any Name? The Values Construct. *Personality and Social Psychology Review* 4, 3 (2000), 255–277. https://doi.org/10.1207/S15327957PSPR0403_4

[35] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40, 2 (nov 2000), 99–-121. https://doi.org/10.1023/A:1026543900054

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

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

[38] Shalom H. Schwartz and Wolfgang Bilsky. 1987. Toward A Universal Psychological Structure of Human Values. *J Pers Soc Psychol* 53 (1987), 550–562.

[39] Marc Serramia, Maite López-Sánchez, and Juan A. Rodríguez-Aguilar. [n.d.]. A Qualitative Approach to Composing Value-Aligned Norm Systems. In *Proceedings of the 19th on Autonomous Agents and Multiagent Systems, AAMAS '20*, Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (Eds.). IFAMAS, 1233–1241.

[40] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. 2021. Value alignment: a formal approach. *CoRR* abs/2110.09240 (2021). arXiv:2110.09240 https://arxiv.org/abs/2110.09240

[41] Colin Stirling. 2001. *Modal and Temporal Properties of Processes*. Springer Verlag.

[42] UNESCO. 2020. Outcome document: first draft of the Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000373434_eng.

[43] Ibo van de Poel. 2018. Design for value change. *Ethics and Information Technology* 23 (2018), 27–-31. https://doi.org/10.1007/s10676-018-9461-9

[44] T. L. van der Weide, F. Dignum, J. J. Ch. Meyer, H. Prakken, and G. A. W. Vreeswijk. 2010. Practical Reasoning Using Values. In *Argumentation in Multi-Agent Systems*, Peter McBurney, Iyad Rahwan, Simon Parsons, and Nicolas Maudet (Eds.). Springer, Berlin, Heidelberg, 79–93.

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