Integrating quantitative and qualitative reasoning for value alignment⋆

Jazon Szabo[1], Jose M. Such[1], Natalia Criado[2], and Sanjay Modgil[1]

[1] King's College London, London, UK
{jazon.szabo,jose.such,sanjay.modgil}@kcl.ac.uk
[2] Universitat Politecnica de Valencia, Spain
ncriado@upv.es

Abstract. Agents that focus only on achieving their own goals may cause significant harm to society. As a result, when deciding which actions to perform, agents have to consider societal values and how their actions impact these values — the 'value alignment problem'. There is therefore a need to integrate quantitative machine reasoning with an ability to reason about the qualitative aspects of human values. In this paper, we present a novel framework for value-based reasoning that aims to bridge the gap between these two modes of reasoning. In particular, our framework extends the theory of grading to model how societal values can trade off with each other or with the agent's goals. Furthermore, our framework introduces the use of hyperreal numbers to represent both quantitative and qualitative aspects of reasoning and help address the value alignment problem.

Keywords: value alignment · practical reasoning · value based reasoning

1 Introduction

Context. The creation of autonomous, intelligent and societally beneficial agents is one of the chief aims of artificial intelligence research. To achieve this, we must create agents that can intelligently reason about what they ought to do, i.e. engage in the process of *practical reasoning* [23]. However, it is not enough for such agents to achieve their own goals; agents that ignore their societal impact may inadvertently cause significant societal harm. This problem – ensuring that the actions of autonomous agents are beneficial to society – is called the *value alignment problem* (*VAP*) [6].

What is beneficial to society (what is considered good, bad, etc) is grounded in the *values* the society upholds. Agents that take into account these values in practical reasoning, are said to engage in *value-based reasoning* [1], wherein encoding information about the relevant values typically requires human input. For example, in a state of the art approach to solving the *VAP* – *cooperative inverse reinforcement learning* (CIRL) [14] – the AI agent gradually learns human preferences (i.e. human values) encoded in a utility function, and simultaneously acts so as to respect the thus-far learnt preferences. The learning process not only involves passive observation of human behaviours, but

⋆ This work was supported by UKRI (EP/S023356/1) and EPSRC (EP/R033188/1).

interactive communication, whereby the human teaches and instructs the AI agent, and the AI agent questions the human. The effectiveness of such a strategy requires that the agent be able to represent the human reasoning communicated to them and incorporate it with their own more quantitatively orientated reasoning. On the other hand, in the so-called *debate game* [15] – also proposed to address the *VAP* – two AI agents exchange instrumental arguments for and against performing certain actions, while evaluative judgment of these arguments is provided by a human judge, thus accounting for human values. As a result, the most value-aligned action is chosen as an outcome of human arbitration. For an agent to be optimal in the debate game, they have to be able to anticipate how the human judge evaluates their arguments, and thus need to accommodate value based reasoning in their deliberations. These two approaches to addressing the *VAP* thus point to the need to integrate human and machine reasoning; that is to say, 'to bridge the gap' between the more fine-grained quantitative reasoning employed by machine learning based AI agents, and the coarser-grained, qualitative aspects of value based reasoning employed by humans.

Existing formalisms and frameworks are either predominantly quantitative or qualitative. Regarding the latter, we are interested in qualitative approaches to value based reasoning; the paradigmatic instance being *value-based argumentation* [2, 4] and other works that build on [2, 4]. In such approaches, actions are judged based on whether they promote or demote a value, from a given set of values that are related by a total order. The main advantage of these qualitative approaches is that they are transparent and understandable. Moreover, qualitative representations are better suited for representing human preferences, especially under uncertainty [22]. However, note that the degree of promotion/demotion, as well as the ordering between the values, is strictly qualitative. As a result, such approaches are too rough-grained and cannot represent quantitative uncertainty or fine-grained trade-offs between various values.

By quantitative approaches, we primarily refer to those assigning each action a utility (e.g. CIRL). Such a representation covers a wide range of approaches, including machine learning, statistical approaches etc. Such methods have historically been very successful and can model fine-grained trade-offs and quantitative uncertainty. However, they are often black-boxes, being difficult to interpret and explain [13]. Moreover, quantitative approaches also cannot accurately represent some ethical values, unlike qualitative approaches [21] (see Section 2.2).

Contributions. In this paper, we bridge the gap between qualitative and quantitative approaches by proposing a framework for value-based reasoning, based on the theory of grading [3] and hyperreal numbers [24]. This framework allows agents to represent and reason both with human values and utility functions. Moreover, the agent can flexibily trade off human values and its utility function.

More precisely, our novel contributions are:

- applying the theory of grading to the problem of value-based reasoning;
- applying the use of hyperreal numbers to the problem of value-based reasoning and to the theory of grading;
- extending the theory of grading with the notion of evaluation facts and show how this relates to the problem of evaluation aggregation;
- extending the theory of grading with the notion of weights;

– proposing a novel framework for value-based reasoning that incorporates the above contributions.

1.1 Motivating example

We now present an example that demonstrates the need for a formalism that bridges the gap between quantitative and qualitative approaches. Moreover, we reuse this example in later sections to showcase how an agent uses our formalism.

In a fictional near future a self-driving boat agent has two aims: 1) to learn to drive at a speed that maximises its task-specific utility 2) to use value-based reasoning to ensure value alignment. The boat can drive at any speed from 0 km/h to 100 km/h. The agent's learnt utility function represents various non explicitly value-based considerations, such as speediness, fuel consumption and comfort. We assume that the expected utility is at a maximum when the speed is 75 km/h and gradually reduces the further the agent deviates from this speed.

Moreover, there are other explicitly value-based considerations distinct from the task utility. At speeds higher than 70 km/h, the ship sends vibrations that affect sea animal behaviour by confusing their senses and thus possibly causing their death. Consider that the fish close to the ship belong to a culturally significant, endangered species. In this case, we expect the agent to first minimise the chance of these animals dying and only optimise the task utility as a secondary consideration.

Now, consider a further complication: the passenger of the boat is stung by a jellyfish. While not in danger, the passenger is in great pain and suffers from a painful fit exactly every sixty minutes. Therefore, to minimise human suffering, the boat needs to carry this passenger to land as fast as possible. However the greater the speed, the more likelihood of causing harm to the fish.

It is not trivial matter to arbitrate as to which is worse: increasing the pain of a human passenger or increasing the probability of killing endangered fish. Moreover, it is not clear how to factor for the qualitative priorities of the different values. Also, how can the agent trade-off these different qualitative priorities with the finer-grained impacts of the actions? And how does the purely quantitative utility function factor? In the remainder of this paper we seek to answer these questions.

2 Requirements

Next we derive specific requirements for bridging the gap between fine-grained quantitative reasoning and the more coarse-grained value-based reasoning.

2.1 Quantitative reasoning

Utility functions are often the end result of intelligent deliberation by an agent. In particular, in many problem domains, if some reasonable axioms (i.e. the von Neumann–Morgenstern axioms) are satisfied, then the preferences of a rational agent can be represented by a utility function [29]. As a result, machine learning agents and many other kinds of rational AI agents are typically modelled as optimising a utility function.

Consequently, we require that 1) the agent be able to use information given in the form of a utility function and 2) the agent use this information in decision making. That is, we require utility compatibility.

2.2 Qualitative reasoning

In value-based reasoning, the standard assumption is that there are multiple values. There are two justifications for this: 1) values are distinct and cannot be reduced to a single value, 2) values may not be distinct but reducing them to a single value is not feasible in practice [9]. As a result, we require that values are modelled separately and that value-based reasoning is formalised as a multi-criteria decision problem.

Furthermore, in ethical decision making, we humans use a more qualitative value-based reasoning [5]. Human understanding of what is right and what is wrong is often abstract and vague [25]. As a result, ethical reasoning cannot be fully represented by the certainty of a total ordering [12]. Consequently, we require that the end product of value-based reasoning is a partial order over the alternatives[3].

Another characteristic of value-based reasoning is that it can be non-Archimedean; we humans often make such evaluative judgments in our decision making. In this context non-Archimedean means a qualitative relationship between two alternatives, where in *any* situation one alternative is preferred to another, regardless of any quantitative considerations. For example, a commonly held deontological principle is that the loss of a human life is a qualitative order of magnitude worse than the loss of money[4]. In particular, this means that any situation where a human loses their life is always worse than any other situation where a human loses any amount money (but not their life). Such qualitative orders of magnitude can be represented only by non-Archimedean quantities. Therefore, we require that the formalism represent non-Archimedean quantities so that we can represent value-based reasoning fully.

2.3 Combining quantitative and qualitative reasoning

Inspired by arguments from Bostrom [7] and Peterson [21], we use hyperreal numbers – a non-Archimedean extension of real numbers – in our proposed framework. In particular: 1) hyperreal numbers can represent any arithmetic operation possible on real numbers[5] and hence are compatible with utility functions; 2) hyperreal numbers can represent non-Archimedean aspects of value-based reasoning that real numbers cannot; and 3) hyperreal numbers can represent the interplay between non-Archimedean quantities and real-valued numbers.

Furthermore, we use a grading-based model in our proposed framework. Grades are a way to model preferences with a degree of 'goodness'. This allows us to model the

[3] Note that this is a generalisation of the standard representation of value-based argumentation, where the total order over values, combined with the strong assumption that each action promotes exactly one value, yields a total order over the actions [2]. By requiring only a partial order over the values, we can relax these underlying strong assumptions.

[4] As is characteristic of many non-consequentialist ethical theories [21].

[5] This is because of the transfer principle. See Section 3.

degree of promotion/demotion of values by different actions. Furthermore, in grading, the amount of information that we have is flexible: it may be more qualitative or more quantitative. Finally, grade aggregation – the process of how alternatives are ranked – takes multiple criteria into account and returns a partial order.

3 Preliminaries

3.1 Hyperreal numbers

Hyperreal numbers [24, 16] are an extension of real numbers containing infinite and infinitesimal quantities. The infinite number ω is such that it is greater than any real number, no matter how large the real number may be, i.e. $\omega > 1 + ... + 1$, for any finite number of terms. The infinitesimal number ϵ is defined as $1/\omega$ and is infinitely close to 0 (but does not equal 0). That is, for any positive real number c, $0 < \epsilon < c$.

We can build the hyperreal numbers via the following rules:

- any real number is a hyperreal number;
- ω and ϵ are hyperreal numbers;
- the sum of any hyperreal numbers is a hyperreal number;
- the product of any hyperreal numbers is a hyperreal number;
- the dividend of any hyperreal numbers is a hyperreal number (but division by zero is not allowed).

We denote the set of hyperreal numbers by $\mathbb{R}^*$.

One of the useful properties of the hyperreal numbers is the *transfer principle* that states that any first-order statements about real numbers also apply to the hyperreal numbers[6]. For example, because multiplication is commutative for the real numbers (e.g., for any reals x, y, it is always true that $xy = yx$), multiplication is also commutative for hyperreal numbers (for any hyperreals x, y, it is always true that $xy = yx$). Because of this, hyperreal numbers behave similarly to real numbers in many ways.

A hyperreal number may be infinitesimal, infinite or finite. A hyperreal number is *infinitesimal* when its absolute value is less than any positive real number. A hyperreal number is *infinite* when its absolute value is greater than any positive real number. Finally, a hyperreal number is *finite* when it is not infinite, that is, when its absolute value is between any two positive real numbers. From this it follows that any infinitesimal number is also finite.

The following properties hold, where c, d are any (non-infinitesimal) finite hyperreals, $\gamma, \delta \neq 0$ are any infinitesimals and N, O are any infinite hyperreals:

- when adding numbers, finite numbers dominate infinitesimal numbers and infinite numbers dominate finite numbers: $\gamma + \delta$ is infinitesimal, $c + \gamma$ is finite, but not infinitesimal, $c + d$ is finite, possibly infinitesimal, and $N + \gamma$, $N + c$ and $N + O$ are both infinite;
- when multiplying numbers, infinitesimal numbers dominate finite numbers and infinite numbers dominate finite numbers: $\gamma \times \delta$ and $\gamma \times c$ are infinitesimal, $c \times d$ is finite but not infinitesimal and $N \times c$ and $N \times O$ are infinite.

[6] However, second-order statements may not transfer to the hyperreals in the same way. For example, the second order statement that there is no number x such that $1 + 1 + 1... < x$ doesn't carry over to the hyperreals.

In the cases we have not given above, the result is not determined: it may be possibly infinitesimal, finite or infinite.

Finally, denote by $a \ll b$ to mean that $a < b$ and a/b is infinitesimal, to be read a is an order of magnitude smaller than b. For example, $\epsilon \ll 1 \ll \omega \ll \omega^2$. Symmetrically we read $a \gg b$ as a is an order of magnitude larger than b.

3.2 Grades

Grades [3, 19] allow to rank alternatives based on their perceived degree of 'goodness'. More precisely, in the theory of grading, different *judges* express their preferences by assigning each alternative a *grade*, where each grade corresponds to a certain standard.

Grades carry some degree of cardinal information. Consequently, grade aggregation can avoid the negative impacts of Arrow's results that affect other preference aggregation methods [17, 3]. However, unlike utility functions, grades often do not carry complete cardinal information, i.e. we do not always know how much better one alternative is than another. How much information is lost by grading depends on the context.

We will now formally define the process of *grade aggregation*, the process of how a set of judges decide out of two alternatives, which one is better, based only on the grades they have given. First, we have a set of social alternatives X judged by a set of judges $N = \{1, ..., n\}$. These judges assign each alternative a grade from a set of grades G, where the set of grades G is ordered by the partial order $\preceq_G$.

The opinions of the judges are represented in their appraisals. That is, the *individual appraisal* of a judge i is denoted by $j_i : X \to G$, a function mapping to each social alternative a grade. As a result, the appraisal of alternative x by judge i is denoted by $j_i(x)$. Moreover, the opinions of the judges together is called their appraisal profile. In more technical detail, an *appraisal profile* J is a list $\langle j_1, ..., j_n \rangle$ of individual appraisals of the judges.

Not every combination of grades is possible. The *appraisal domain* $\mathcal{D}$ is the set of appraisal profiles that the judges might produce for a specific grading scenario. When aggregating grades, the grading rules only have to specify what to do for appraisal profiles in the appraisal domain, where a *grading rule* is a function f such that f maps any appraisal profile $J \in \mathcal{D}$ in the grade domain to a partial binary relation $f(J) = \preceq$ on the set of social alternatives X. Intuitively, a grading rule takes the grades assigned to each individual by the different judges and produces a ranking over them.

4 Grade-based framework of value-based reasoning

In this section, we introduce a framework that satisfies the previously identified requirements in order to bridge the gap between quantitative and qualitative value-based reasoning. To do this, we base our framework on the theory of grading and extend it with a notion of grade information, judge weights and introduce the use of hyperreal numbers to this context.

In grading, a set of *judges appraise* the *social alternatives* in order to rank them from best to worst. Similarly, in value-based reasoning, a set of *values evaluate* the possible *actions* of the agent in order to rank them from best to worst. Therefore, the judges of value-based reasoning are the values and the social alternatives are the actions.

More precisely, in value-based reasoning, the possible actions A are evaluated by the agent's values V. Various values can be promoted and demoted to different levels; the set of all such levels of evaluation is denoted by L and is ordered by a partial order $\preceq_L$. The levels of evaluation in value-based reasoning correspond to the grades of grading. The *individual evaluation* of an action a with respect to a value v corresponds to what level of value promotion/demotion performing action a brings about.

Definition 1 (Individual evaluation). *Given an agent, a value v that the agent holds and a set of actions A that the agent may perform, the individual evaluation $e_v : A \to L$ is a function that maps any action a to the corresponding level of evaluation.*

Example. Driving at a speed of 55 km/h means that no animals are harmed. This is the status quo, which we denote with the neutral evaluation 0 (zero). Therefore, we denote this individual evaluation as $e_a(drive(55)) = 0$, where $drive(55)$ is the action of driving the boat with a speed of 55 km/h and a is the value of animal welfare.

The *evaluation profile* E of an agent is a list of the individual evaluations of each of the agent's values $E = \langle e_{v_1}, ..., e_{v_{|V|}} \rangle$. The *evaluation domain* D is a set of evaluation profiles that the agent's values may produce for a specific scenario. The agent uses an *evaluation aggregation rule* to order their actions based on the evaluations of the different values.

Definition 2 (Evaluation aggregation rule). *Given an evaluation domain D and a set of actions A, the evaluation aggregation rule is a function $f : D \to PO(A)$ that maps any evaluation profile $E \in D$ in the evaluation domain to a partial order over the set of actions $f(E) = \preceq_A$, where $PO(A)$ is the set of all partial orders over A.*

Example. An evaluation aggregation rule tells the agent how to trade off the different levels of promotion/demotion for the various values of the agent. Consider a simplified version of the example: there are only two values, animal welfare a and utility u; there are only two actions, $drive(55)$ and $drive(75)$. Furthermore, assume that we know that driving at a speed of 75 km/h is expected to kill precisely one endangered fish. Moreover, assume $u(drive(55)) = \frac{11}{15}$ and $u(drive(75)) = 1$. Hence, for the actions $drive(55)$ and $drive(75)$ we have:

- $e_a(drive(55)) = 0$ and
- $e_u(drive(55)) = \frac{11}{15}$.
- $e_a(drive(75)) = d$ and
- $e_u(drive(75)) = 1$,

where d is the level of demotion ($d < 0$) that occurs when an endangered fish dies.

Take an evaluation aggregation rule f such that the resulting order $\preceq_A$ contains $drive(75) \prec_A drive(55)$. This rule f implies that increasing the level of promotion of the value of utility from $\frac{11}{15}$ to 1 is strictly worse than demoting the value of animal welfare with a level d. That is, the evaluation aggregation rule f states that the increase in the task utility is not worth the decrease in animal welfare.

Finally, note that the evaluation domain can be used to denote that different values are evaluated differently. For example, by requiring in the evaluation domain that only those evaluation profiles that map every utility function to a real number are permitted, we denote that the value of utility is always evaluated as a real number.

4.1 Evaluation aggregation rules

Evaluations by the values correspond to how good the different actions are. That is, they carry some cardinal information, which can be used when comparing the different actions. Therefore, the evaluations cannot be aggregated in arbitrary ways; the evaluations should only be aggregated using evaluation aggregation rules that are compatible with the information they carry. Therefore, we define this notion of *compatibility* and apply it in the context of value-based reasoning.

First, we define the information given by the evaluations of values in the form of equations and inequalities. Such equations and inequalities are called the *evaluation information*. That is, the levels of evaluation correspond to hyperreal variables where the information that we know about them restrict the possible range of hyperreal values they may take.

An *aggregation fact* describes how two different sets of evaluations should be ordered, regardless of what values the evaluations belong to.

Definition 3 (Aggregation fact). *An aggregation fact is a tuple (L_1, L_2, op) where L_1 and L_2 are multisets of levels of evaluation and $op \in \{<, \leq, =, \ll\}$ is a comparison operation. Such an aggregation fact is interpreted as the statement $(\Sigma_{l \in L_1} l)$ op $(\Sigma_{l' \in L_2} l')$.*

Example. We can formalise with an aggregation fact the requirement that increasing the level of promotion of *any* value from $\frac{11}{15}$ to 1 is strictly worse than demoting *any* other value with a level d. This corresponds to the aggregation fact $(\{d, 1\}, \{0, \frac{11}{15}\}, <)$, which we interpret as the inequality $(d + 1) < (0 + \frac{11}{15})$.

The collection of all aggregation facts for a set of values is named *aggregation information*. Formally, the aggregation information I is a set of aggregation facts.

We further require for simplicity's sake, that the partial order $\preceq_L$ over the levels of evaluation be encompassed within the aggregation information. This is to ensure that the aggregation information contains all relevant information about how the different levels of information are related to one another. More precisely, if for some values a and b, $a \preceq_L b$ holds, then the aggregation information has to contain the corresponding aggregation fact, i.e. $(\{a\}, \{b\}, \leq) \in I$.

4.2 Using the aggregation information

We now define how the aggregation information can be used to aggregate the evaluations. First, note that a collection of aggregation facts forms a system of equations and inequalities. Here, the variables are the levels of evaluation of the values. Consequently, the evaluation information defines for each level of evaluation a range of possible hyperreal values that is compatible with the information given.

An assignment to each of the level variables is called a *satisfying assignment* if the system of equations and inequalities created by the set of aggregation information is satisfied by the substituted values. Note that for a single variable, there may be a range of different assignments possible that are consistent with the information.

Definition 4 (Assignment (unweighted)). *Given a set of levels of evaluation L, a function is an assignment function $as : L \rightarrow \mathbb{R}^*$ if it maps every level of evaluation $l \in L$ to a hyperreal number $as(l)$.*

Note that we sometimes abuse notation and use the same symbols for levels of evaluation and their assignments interchangeably. For example, utility functions' levels of evaluation are denoted by numbers and not as variables. Similarly, we use the number 0 to refer to the level of neutral evaluation and 1 to the level of unit promotion.

We now define the notion of a satisfying assignment by first defining it for an aggregation fact and then for aggregation information.

Definition 5 (Satisfying assignment — fact (unweighted)). *Given a set of levels of evaluation L, an aggregation fact $af = (L_1, L_2, op)$, an assignment function $as : L \rightarrow \mathbb{R}^*$ satisfies the aggregation fact af if the substituted interpretation $(\Sigma_{l \in L_1} as(l))$ op $(\Sigma_{l' \in L_2} as(l'))$ is true.*

Definition 6 (Satisfying assignment — information (unweighted)). *Given a set of levels of evaluation L, a set of aggregation information I, an assignment function $as : L \rightarrow \mathbb{R}^*$ satisfies the information I, if for every fact $af \in I$ in the information, as satisfies af.*

Example. Consider the fact $(\{d, 1\}, \{0, \frac{11}{15}\}, <)$. This fact is satisfied by the function as, where $as(d) = -1$. This is because the interpretation $(as(d) + 1) < (0 + \frac{11}{15})$ is satisfied, as $-1 + 1 = 0 < \frac{11}{15}$. In fact, we can see that any assignment that maps d to a value less than $-\frac{4}{15}$ will be a satisfying assignment.

Aggregation information may have multiple satisfying assignments. The set of all such assignments is called the *solution set* for the information.

Definition 7 (Solution set (unweighted)). *Given aggregation information I, the solution set S_I is the set such that function as is in the solution set $as \in S_I$ iff the function as satisfies the information I, i.e. $S_I = \{as : L \rightarrow \mathbb{R}^* | as \text{ satisfies } I\}$.*

Example. Consider the information $I = \{(\{d, 1\}, \{0, \frac{11}{15}\}, <)\}$. Remember that this is satisfied by any function as such that d is assigned less than $-\frac{4}{15}$, i.e. $as(d) < -\frac{4}{15}$. Therefore, the solution set is given as $S_I = \{as : L \rightarrow \mathbb{R}^* | as(d) < -\frac{4}{15}\}$.

4.3 Weighing the values

Recall that the aggregation facts presented in the previous section do not distinguish between different values. This is not always appropriate; the agent may prioritise different values to different degrees. These weighted priorities affect how different values and their respective levels of evaluations are compared.

We model these priorities through a set of weights W, where the weights are hyperreal variables. A *weight profile* maps each value v to a weight $w(v) \in W$.

Definition 8 (Weight profile). *Given a set of values V and a set of weights W, a weight profile is a function $w : V \to W$ such that $w(v) \in W$ denotes the weight of value v.*

Example. The agent has three values: utility u, animal welfare a and human welfare h. Consider that the agent prioritises human welfare over animal welfare and utility in every situation. This can be represented by the weight profile w, that is, $w(u) = 1$, $w(a) = 1$ and $w(h) = \omega$.

Through the use of weights, we can declare how different levels of evaluation of specific values should trade off. We specify such through *weighted aggregation facts* and *weighted aggregation information*, which generalise aggregation facts and aggregation information, respectively.

Definition 9 (Weighted aggregation fact). *Given a set of weights W, levels of evaluation L, a weighted aggregation fact is a tuple (WL_1, WL_2, op) where WL_1 and WL_2 are multisets of pairs (w, l), $w \in W$ and $l \in L$, and $op \in \{<, \leq, =, \ll\}$ is a comparison operation. A weighted aggregation fact is interpreted as the statement $(\Sigma_{(w,l)\in WL_1} wl)$ op $(\Sigma_{(w',l')\in WL_2} w'l')$.*

The collection of all weighted aggregation facts for a set of values is called the *weighted aggregation information*.

Example. Let $w(u)$ represent the weight of the utility function and $w(a)$ represent the weight of animal welfare. We can now formalise with a weighted aggregation fact a requirement from a previous example; namely, that increasing the level of promotion of the value of utility from $\frac{11}{15}$ to 1 is strictly worse than demoting the value of animal welfare to d. This corresponds to the weighted aggregation fact $(\{(w(a), d), (w(u), 1)\}, \{(w(a), 0), (w(u), \frac{11}{15})\}, <)$, which we interpret as the inequality $(w(a)d + w(u)1) < (w(a)0 + w(u)\frac{11}{15})$.

Note. Weighted aggregation facts are strictly more expressive than (unweighted) aggregation facts. This is because unweighted aggregation facts can be understood as uniformly weighted and so can be expressed as weighted aggregation facts. For example, if $1 \in W$, then we can express the (unweighted) aggregation fact (L_1, L_2, op) as weighted aggregation fact (WL_1, WL_2, op), where $WL_1 = \{(1, l) | l \in L_1\}$ and $WL_2 = \{(1, l) | l \in L_2\}$.

Note. We can express a partial order $\preceq_W$ over the weights by a set of facts. In particular, if $1 \in L$ we can express $w_1 \preceq_W w_2$ as the fact $(\{(w_1, 1)\}, \{(w_2, 1)\}, \leq)$. By including all such facts, we can express any partial order $\preceq_W$ in the weighted aggregation information.

Example. If we care about animal welfare a and utility u to the same degree, we can express this as $(\{(w(a), 1)\}, \{(w(u), 1)\}, =)$, that is, $w(a) = w(u)$.

Furthermore, we can extend the definition of a satisfying assignment to apply to weighted aggregation facts and information. Note that to be able to interpret the weighted facts, we have to assign the weights a specific hyperreal value as well.

Definition 10 (Assignment (weighted)). *Given a set of levels of evaluation L and a set of weights, a function is an assignment function $as : (L \cup W) \to \mathbb{R}^*$ if it maps every level of evaluation $l \in L$ to a hyperreal number $as(l)$ and every weight $w \in W$ to a hyperreal number $as(w)$.*

Note that while it is possible to assign a value a non-positive weight[7], we assume that if an agent holds a value v then the agent seeks to achieve that value, i.e. the agent weights the value positively, i.e. for all v $as(w_v) > 0$ holds.

Definition 11 (Satisfying assignment — fact (weighted)). *Given a set of levels of evaluation L, a weighted aggregation fact $waf = (WL_1, WL_2, op)$, an assignment function $as : (L \cup W) \to \mathbb{R}^*$ satisfies the weighted aggregation fact waf if the substituted interpretation $(\Sigma_{(w,l) \in WL_1} as(w)as(l))$ op $(\Sigma_{(w',l') \in WL_2} as(w')as(l'))$ is true.*

Definition 12 (Satisfying assignment — information (weighted)). *Given a set of levels of evaluation L, a set of weights W, a set of weighted aggregation information I, an assignment function $as : (L \cup W) \to \mathbb{R}^*$ satisfies the information I, if for every fact $waf \in I$ in the information, as satisfies waf.*

Example. Consider the previously identified weighted aggregation fact $(\{(w(a), d), (w(u), 1)\}, \{(w(a), 0), (w(u), \frac{11}{15})\}, <)$. This fact is interpreted as $(w(a)d + w(u)1) < (w(a)0 + w(u)\frac{11}{15}) = w(u)\frac{11}{15}$. Therefore, we have $w(a)d < -w(u)\frac{4}{15}$. Let as be an assignment such that $as(w(a)) = 1$, $as(w(u)) = \omega$ and $as(d) = -\omega^2$. In this case, both the left hand side and the right hand side are negative infinities. However, the absolute value of the left hand side is an order of magnitude larger (ω^2 compared to ω) and so the left hand side is less than the right hand side. As a result, as is a satisfying assignment.

Aggregation information may have multiple satisfying assignments. The set of all such assignments is called the solution set for the information.

Definition 13 (Solution set (weighted)). *Given weighted aggregation information I, the solution set S_I is the set such that function as is in the solution set $as \in S_I$ iff the function as satisfies the information I, i.e. $S_I = \{as : (L \cup W) \to \mathbb{R}^* | as \text{ satisfies } I\}$.*

Example. Consider the previously identified weighted aggregation fact $(\{(w(a), d), (w(u), 1)\}, \{(w(a), 0), (w(u), \frac{11}{15})\}, <)$. This fact is interpreted as $(w(a)d + w(u)1) < (w(a)0 + w(u)\frac{11}{15}) = w(u)\frac{11}{15}$. Therefore, we have $w(a)d < -w(u)\frac{4}{15}$. Also consider that the agent weights these two values equally, i.e. the fact $(\{(w(a), 1)\}, \{(w(u), 1)\}, =)$ holds. As a result, we have $w(a) = w(u)$. From our assumption that every weight is positive, we can then divide by $w(a)$ and conclude that $d < -\frac{4}{15}$. Therefore, the solution set can be defined as $S_I = \{as : (L \cup W) \to \mathbb{R}^* | as(d) < -\frac{4}{15} \wedge as(w(a)) = as(w(u))\}$.

Note. To simplify matters, we assume that the information that the agent is given is consistent, i.e. has always at least one possible solution. In practice, however, if the information is inconsistent, then the solution set will be empty.

[7] This can be used to model values the agent is apathetic towards (zero weight) or is even hostile to (negative weight).

4.4 Compatibility

We have thus far defined what assignment of variables are consistent with given information. We now define how an evaluation aggregation rule can be compatible with the information given, based on the possible assignment of variables.

In particular, we are interested in ranking the actions through the evaluation aggregation rule based on the information available and nothing else. That is, if one alternative is better than another under all satisfying assignments, then it must be ordered as the better action. On the other hand, if for different satisfying assignments, we may order the actions differently, then the aggregation rule should consider the two actions to be incomparable.

Definition 14 (Compatibility). *Given weighted evaluation information I, an evaluation profile $E = \langle e_{v_1}, ..., e_{v_{|V|}} \rangle$ in the appraisal domain D, values V, weight profile w and actions A, an evaluation aggregation rule f is compatible with weighted evaluation information I, if for any actions $a, b \in A$ the statement $a \preceq_A b$ holds iff for all satisfying assignments $as \in S_I$ in the solution set it holds that $\Sigma_{v \in V}\, as(w(v))as(e_v(a)) \leq \Sigma_{v' \in V}\, as(w(v'))as(e_{v'}(b))$, where $\preceq_A = f(E)$.*

Note. Since compatibility uniquely determines whether for any actions a, b, $a \preceq_A b$ holds or not, there is at most one compatible evaluation aggregation rule. If the information is consistent, then we also know there is at least one compatible evaluation aggregation rule. Therefore, for consistent weighted aggregation information, we have a unique compatible rule. We denote this rule by f_I.

This means that if in our framework we formalise a value-based reasoning problem fully and appropriately, then there is a unique evaluation aggregation rule that is compatible with the evaluation information. This means that if the agent orders their action based on f_I then any of the maximal elements of the resulting partial order is guaranteed to be a value-aligned action.

Example. Consider the facts $(\{(w(a), 1)\}, \{(w(u), 1)\}, =)$ and $(\{(w(a), d), (w(u), 1)\}, \{(w(a), 0), (w(u), \frac{11}{15})\}, <)$. From these, we derived the solution set $S_I = \{as : (L \cup W) \to \mathbb{R}^* | as(d) < -\frac{4}{15} \wedge as(w(a)) = as(w(u))\}$.

Now consider the actions $drive(55)$ and $drive(75)$ and their evaluations E by the values utility u and animal welfare a:

- $e_a(drive(55)) = 0$ and
- $e_u(drive(55)) = \frac{11}{15}$.
- $e_a(drive(75)) = d$ and
- $e_u(drive(75)) = 1$,

where d corresponds to the demotion caused by an endangered fish dying.

The weighted sum of evaluations, using any satisfying assignment for driving at a speed of 55 km/h is (which we denote by sum_{55}) $sum_{55} = as(w(u))as(e_u(drive(55)))+ as(w(a))as(e_a(drive(55))$. Using the above, we derive that $sum_{55} = as(w(u))\frac{11}{15}$.

The weighted sum of evaluations for driving at a speed of 75 km/h is $sum_{75} = as(w(u))as(e_u(drive(75))+as(w(a))as(e_a(drive(75)) = as(w(u))+as(w(a))as(d)$. Since for all as, $as(w(u)) = as(w(a))$, we have $sum_{75} = as(w(u))(1 + as(d))$.

We also know that $as(d) < -\frac{4}{15}$ and so $sum_{75} < as(w(u))(\frac{11}{15})) = sum_{55}$.

Therefore, for all as, $sum_{75} < sum_{55}$, so the unique compatible aggregation evaluation rule f_I aggregates the evaluations E such that $drive(75) \prec_A drive(55)$, where $\preceq_A = f_I(E)$.

4.5 Applying formalism to example

We now showcase how an agent could utilise our proposed formalism, using our motivating example from section 1.1. We assume that the agent has three distinct values: the task utility (u), animal welfare (a) and human welfare (h). We also assume that the agent prioritises human welfare over animal welfare and utility. That is, the partial order over the weights of the values is given as $w(u), w(a) \ll w(h)$.

This agent chooses the action that maximises the total level of evaluation of the different values. To handle uncertainty, the agent uses expectations to calculate the sum of the evaluations. For the 'value' of utility, it's simply the expected utility, which is given by the function $u(x)$. For simplicity's sake, we assume that the utility function u is given by the following equation:

$$u(x) = \begin{cases} x/75 & x \leq 75 \\ (100 - x)/25 & x > 75 \end{cases}$$

where x is the speed of the boat in km/h. We can see that the expected utility is maximal when the speed is 75 km/h and gradually reduces the further the speed deviates from 75.

For animal welfare a, we assume the expectation is:

$$e_a(drive(x)) = \begin{cases} 0 & 70 \geq x \\ 1d & 80 \geq x > 70 \\ 2d & 90 \geq x > 80 \\ 3d & 100 \geq x > 90 \end{cases}$$

Here, d is the level of demotion associated with the death of a single endangered fish. Furthermore, assume that $d \ll 0$; that is, the death of a single fish is a level of magnitude worse than the status quo. Consequently, this value is maximally promoted when the agent maintains the status quo by driving slower than 70 km/h.

For human welfare, recall that the passenger is in pain after being stung by a jellyfish. This pain is not distributed evenly: the passenger gets painful fits after exactly sixty minutes of completed journey. Therefore, the amount of pain depends on the length of the journey. Assume that the journey is 210 km. Thus, the expectation is:

$$e_h(drive(x)) = p\lfloor 210/x \rfloor$$

Here $p < 0$ corresponds to the level of demotion from the pain of a fit. Therefore, this value is maximally promoted when the agent minimises human pain by minimising the number of fits, which happens in the range $(70, 100]$.

We represent the agent's objective, i.e. the sum of the weighed expectations, by the function $V(x)$. More precisely, $V(x) = w(u)u(x) + w(a)e_a(drive(x)) + w(h)e_h(drive(x))$.

We can see that the agent's different values evaluate the range of possible velocities differently. We can use the weights of the values to find the optimal speed, that is, the speed x which maximises $V(x)$.

Let us start with a simpler case, where $w(a) = 0$ and $w(u) = 1$; that is, the agent is indifferent towards animal welfare but values the learnt utility. In this case, we can see that the agent prioritises maximising human welfare first, and then the learnt utility. This is because $w(u) \ll w(h)$. The agent maximises human welfare by only accepting speeds within the range $(70, 100]$ and then maximises the utility by picking a value from within this range that maximises the utility. This local maximum of the utility function is the global maximum, which is when $x = 75$. Therefore, in this simpler case, the agent would choose to drive at a speed of 75 km/h.

Now, consider the more complex case where $w(a) = w(u) = 1$. First, we see that for speeds $x \in (90, 100]$ we have $V(x) = u(x)+3d+w(h)2p$. For speeds $x \in (80, 90]$ we have $V(x) = u(x) + 2d + w(h)2p$. We can see that this range is better than the previous one, as $d \ll 0$. For speeds $x \in (70, 80]$, we have $V(x) = u(x) + d + w(h)2p$. This is similarly better than the latter range. Therefore, in the range $(70, 100]$, the range $(70, 80]$ is optimal; within this range, the exact speed of 75 km/h maximises the utility. Therefore, $x = 75$ is optimal in this range.

For speeds $x \in (52.5, 70]$, we have $V(x) = u(x) + w(h)3p$. For speeds $x \in (0, 52.5]$, we have $V(x) < u(x) + w(h)4p$, which is worse than the latter range, as $w(h)p \ll 0$. Therefore, in the range $(0, 70]$, the optimal range of speed is in the range $(52.5, 70]$. In this range, the exact speed of 70 km/h maximimises the utility. Therefore, $x = 70$ is optimal in this range.

To derive whether $x = 75$ or $x = 70$ is better, we have to compare $V(75) = u(75) + d + w(h)2p$ and $V(70) = u(70) + w(h)3p$. We can do this by evaluating $V(70) - V(75) = u(70) - u(75) + w(h)p - d$.

Note that $u(70) - u(75) = -1/15$. Moreover, we know that $w(h)p \ll 0$ as $w(h) \gg 0$ and $p < 0$. Further, we also know that $d \ll 0$. From the properties of hyperreal numbers, we know that the sum of a (positive) infinity $-d$ and a (negative) infinity $w(h)p$ may be infinitesimal, finite or infinite. Therefore, based on the information we have, we cannot decide which of the options is better. As a result, there are two maximal elements of the partial order of the possible actions, $drive(75)$ and $drive(70)$.

When there is a tie, it may be because the agent does not have enough information. In such cases, the agent may ask for information from a human. Imagine that the agent learns from communication the weighted evaluation information $(\{1\}, \{d\}, \{w(h)\}, \{p\}, <)$. This is interpreted as $d < w(h)p$. That is, the death of an endangered fish is worse than one fit of pain for one person. Note that this is still not enough to decide between the two alternatives, however, because of the utility. That is, even though we know that $0 < w(h) - d$, we do not know whether $0 < -1/15 + w(h)p - d$.

Now consider instead that the agent learns the information $(\{1\}, \{d\}, \{w(h)\}, \{p\}, \ll)$, that is, $d \ll w(h)p$. Then by the properties of the hyperreal numbers, we would also know that $0 \ll -1/15 + w(h) - d$ and hence that 70 is an optimal speed. In other words, if the agent prioritised the survival of animal species over the lessening of human suffering, the agent would prefer to drive at 70 km/h.

Finally, consider the agent has the information $(\{w(h)\}, \{p\}, \{1\}, \{d\}, \leq)$, which is interpreted as $w(h)p \leq d$. That is, one painful fit is not better than the death of an endangered fish. In this case, this information is enough to decide, because $0 \leq d - w(h)p$ and so it must also be that $0 \leq w(h)p - d$ and, moreover, that $0 > -1/15 + w(h)p - d$ holds. That is, if the agent believes that the painful fit is at least as bad as the death of an endangered fish, the agent would prefer to drive at 75 km/h.

5 Related Work

In this section, we divide the related work into two broad categories: whether the values are represented in a more qualitative or a more quantitative way.

5.1 Related Qualitative Approaches

Argumentation emerged as an alternative approach to classical utility-based decision theory, where argument and logic-based representations are the basis for decision making [31]. Which argument is accepted is subjective and depends on the *audience* (effectively a total ordering on values). As a result, subjective human preferences, i.e. values, can model which argument succeeds and so what decision should be made, based on a specific audience [2]. This is modelled in *value-based argumentation* [2, 4], where agents propose arguments in favour of different actions. Note that each argument promotes exactly one value and there is a total order over the values (i.e. the audience's preferences). The different arguments may attack one another; an attack is successful if the defending argument's value does not promote a more important value than the attacking argument's value. However, value-based argumentation is limited in expressiveness: each argument can promote only one value and no other level of evaluation is possible, (i.e. no demotion, no magnitude of promotion etc.). Moreover, arguments are decided solely based on the total ordering of values, i.e. no fine-grained trade-offs between the different arguments are possible.

To remedy this, many subsequent models have used more expressive representations of evaluations. For example, Serramia et al. [26] allow values to label alternatives with a degree of promotion or demotion. Then, the alternatives are ordered based on which values promote and demote different alternatives and to what levels. This approach also considers how many values of the same importance are promoted and this is traded-off against how many values are demoted. Therefore, some basic quantitative considerations can be represented in this approach. However, values are only counted against other values from the same rank (i.e. importance) and are all uniformly weighed within the same rank. Furthermore, the rank of a value and the level of evaluation are rigid and cannot be flexibly traded-off.

Zurek and Mokkas propose a framework [30, 31], which allows for a more flexible trade-off between the various values and their different levels of promotion/demotion. They do this by prescribing only some basic properties of how individual evaluations should be aggregated. Similarly to our framework, their approach also permits the input of agent knowledge about how different levels of evaluation and value importance are related. However, their framework does not allow for the representation of anti-Archimedean priorities over the importance of different values or levels of evaluation.

5.2 Related Quantitative Approaches

Similarly to the qualitative methods, many quantitative methods associate with alternatives a degree of promotion/demotion; such approaches then use the weighted sum of values and their degrees of promotion and demotion [11, 27, 18, 28]. The advantage of this is the flexible and fine-grained trade-off of values and their levels of promotion and demotion. For example Serramia et al. [27] use the sum of the promoted values's weights and compare it against the relative cost. On the other hand, Szabo et al. [28] use the weighted sum of evaluations to filter out desires and intentions of a BDI agent that would cause the agent to violate their own values. However, these approaches cannot represent qualitative priorities between different values.

Values are also often represented through utility functions [8, 14, 20, 29]. For example, human preferences may be learned through cooperative inverse reinforcement learning [14] or other machine learning methods. An agent that acts based on an appropriate utility function would behave in a value-aligned way [20]. However, while utility-based frameworks allow for a more fine-grained trade-off between various values, they do not allow for more qualitative, anti-Archimedean trade-offs between them.

6 Conclusion and future work

Computational agents need to reason as to how to accomplish their goals while accounting for the values of the society in which they are embedded. As a result, there is a need to integrate the primarily quantitative based reasoning distinctive of machine learning based AI with qualitative human values. To address this problem, we have identified various requirements and presented a novel formalism that can integrate various properties of quantitative and qualitative reasoning for value alignment. To do so, our framework uses a combination of hyperreal numbers and a grading-based model.

Moreover, we conjecture that our framework *generalises* quantitative and qualitative reasoning in the context of *VAP*. We further conjecture that generalising the two requires the use of a non-Archimedean extension of real numbers[8], such as the hyperreal numbers. This is because of the utility compatibility and non-Archimedean requirements. Finally, we also argue that evaluations carry different amounts of information (sometimes finer-grained, sometimes rougher-grained) and a grading-based model can capture these features.

We have described how to represent the problem of value-based reasoning through our formalism. While we have described an informal example of how an agent may use this formalism, future work should propose an agent framework that makes use of this formalism, and will require: (i) an algorithm that can efficiently reason about the aggregation information and derive the value aligned evaluation aggregation rule; (ii) an algorithm for learning aggregation facts from observation and communication.

[8] Hyperreal numbers are not the only such number system; for example, surreal numbers [10] are also a non-Archimedean extension of real numbers. We have decided to use hyperreal numbers because of the transfer principle.

References

1. Atkinson, K., Bench-Capon, T.J.M.: Taking account of the actions of others in value-based reasoning. Artif. Intell. **254**, 1–20 (2018). https://doi.org/10.1016/j.artint.2017.09.002, https://doi.org/10.1016/j.artint.2017.09.002
2. Atkinson, K., Bench-Capon, T.J.M.: Value-based argumentation. FLAP **8**(6), 1543–1588 (2021), https://collegepublications.co.uk/ifcolog/?00048
3. Balinski, M., Laraki, R.: A theory of measuring, electing, and ranking. Proceedings of the National Academy of Sciences **104**(21), 8720–8725 (2007)
4. Bench-Capon, T.J.M.: Value-based argumentation frameworks. In: Benferhat, S., Giunchiglia, E. (eds.) 9th International Workshop on Non-Monotonic Reasoning (NMR 2002), April 19-21, Toulouse, France, Proceedings. pp. 443–454 (2002)
5. Bench-Capon, T.J.M., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artif. Intell. Law **25**(1), 29–64 (2017). https://doi.org/10.1007/s10506-017-9194-9, https://doi.org/10.1007/s10506-017-9194-9
6. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Inc., USA, 1st edn. (2014)
7. Bostrom, N., et al.: Infinite ethics. Analysis and Metaphysics (10), 9–59 (2011)
8. Brown, D.S., Schneider, J., Dragan, A.D., Niekum, S.: Value alignment verification. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 1105–1115. PMLR (2021), http://proceedings.mlr.press/v139/brown21a.html
9. Chang, R.: Value pluralism. In: Wright, J. (ed.) International Encyclopedia of the Social and Behavioral Sciences (Second Edition), pp. 21–26. Elsevier (2015)
10. Conway, J.H.: On numbers and games. AK Peters/CRC Press (2000)
11. Cranefield, S., Winikoff, M., Dignum, V., Dignum, F.: No pizza for you: Value-based plan selection in BDI agents. In: Sierra, C. (ed.) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. pp. 178–184. ijcai.org (2017). https://doi.org/10.24963/ijcai.2017/26, https://doi.org/10.24963/ijcai.2017/26
12. Eckersley, P.: Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). In: Espinoza, H., hÉigeartaigh, S.Ó., Huang, X., Hernández-Orallo, J., Castillo-Effen, M. (eds.) Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019. CEUR Workshop Proceedings, vol. 2301. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2301/paper_7.pdf
13. Geffner, H.: Model-free, model-based, and general intelligence. In: Lang, J. (ed.) Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. pp. 10–17. ijcai.org (2018). https://doi.org/10.24963/ijcai.2018/2, https://doi.org/10.24963/ijcai.2018/2
14. Hadfield-Menell, D., Russell, S.J., Abbeel, P., Dragan, A.D.: Cooperative inverse reinforcement learning. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 3909–3917 (2016), https://proceedings.neurips.cc/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html
15. Irving, G., Christiano, P.F., Amodei, D.: AI safety via debate. CoRR **abs/1805.00899** (2018), http://arxiv.org/abs/1805.00899

16. Keisler, H.J.: The hyperreal line. In: Real numbers, generalizations of the reals, and theories of continua, pp. 207–237. Springer (1994)
17. Little, I.M.: Social choice and individual values. Journal of Political Economy **60**(5), 422–432 (1952)
18. Mercuur, R., Dignum, V., Jonker, C.M.: The value of values and norms in social simulation. J. Artif. Soc. Soc. Simul. **22**(1) (2019). https://doi.org/10.18564/jasss.3929, https://doi.org/10.18564/jasss.3929
19. Morreau, M.: Grading in groups. Economics & Philosophy **32**(2), 323–352 (2016)
20. Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Varshney, K.R., Campbell, M., Singh, M., Rossi, F.: Teaching AI agents ethical values using reinforcement learning and policy orchestration. IBM J. Res. Dev. **63**(4/5), 2:1–2:9 (2019). https://doi.org/10.1147/JRD.2019.2940428, https://doi.org/10.1147/JRD.2019.2940428
21. Peterson, M.: A royal road to consequentialism? Ethical theory and moral practice **13**(2), 153–169 (2010)
22. Popova, A., Regenwetter, M., Mattei, N.: A behavioral perspective on social choice. Annals of Mathematics and Artificial Intelligence **68**(1), 5–30 (2013)
23. Raz, J.: Practical reasoning. Oxford University Press (1978)
24. Robinson, A.: Non-standard analysis. Princeton University Press (2016)
25. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 2307–0919 (2012)
26. Serramia, M., López-Sánchez, M., Moretti, S., Rodríguez-Aguilar, J.A.: On the dominant set selection problem and its application to value alignment. Auton. Agents Multi Agent Syst. **35**(2), 42 (2021). https://doi.org/10.1007/s10458-021-09519-5, https://doi.org/10.1007/s10458-021-09519-5
27. Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A., Morales, J., Wooldridge, M.J., Ansótegui, C.: Exploiting moral values to choose the right norms. In: Furman, J., Marchant, G.E., Price, H., Rossi, F. (eds.) Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018. pp. 264–270. ACM (2018). https://doi.org/10.1145/3278721.3278735, https://doi.org/10.1145/3278721.3278735
28. Szabo, J., Such, J.M., Criado, N.: Understanding the role of values and norms in practical reasoning. In: Bassiliades, N., Chalkiadakis, G., de Jonge, D. (eds.) Multi-Agent Systems and Agreement Technologies - 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers. Lecture Notes in Computer Science, vol. 12520, pp. 431–439. Springer (2020). https://doi.org/10.1007/978-3-030-66412-1_27, https://doi.org/10.1007/978-3-030-66412-1_27
29. Von Neumann, J., Morgenstern, O.: Theory of games and economic behavior. Princeton university press (2007)
30. Zurek, T.: Goals, values, and reasoning. Expert Syst. Appl. **71**, 442–456 (2017). https://doi.org/10.1016/j.eswa.2016.11.008, https://doi.org/10.1016/j.eswa.2016.11.008
31. Zurek, T., Mokkas, M.: Value-based reasoning in autonomous agents. Int. J. Comput. Intell. Syst. **14**(1), 896–921 (2021). https://doi.org/10.2991/ijcis.d.210203.001, https://doi.org/10.2991/ijcis.d.210203.001