A Review of AI Agent Reasoning with Values

Jay Paul Morgan[a,*] **and Adam Wyner**[a,**]

[a]Department of Computer Science, Swansea University, Swansea, United Kingdom
ORCID (Jay Paul Morgan): https://orcid.org/0000-0003-3719-362X, ORCID (Adam Wyner): https://orcid.org/0000-0002-2958-3428

Abstract. With the increasing presence and integration of technology in our society, they are more frequently involved in fundamental decisions that have a significant impact on us. It is only natural that, if decisions will be automated by technology, we would wish to guarantee they are optimally good and well-behaved according to ethical (moral, value, deontic) requirements. One approach which seeks to integrate values within the decision process is called Value-based Reasoning. In this study, we conduct a systematic review of the literature on Value-based Reasoning to gain a clear understanding of the broader landscape of this field and to determine its future direction.

1 Introduction

As technology becomes more advanced and capable, it becomes a larger part of our society. Its role in society opens the door for many instances of automated decision-making, which has a (potentially detrimental) effect on people's lives [73]. The use of automated decision-making AI agents, in particular, raises numerous ethical concerns which include, but certainly not limited to, privacy issues in its use of data, the autonomy of humans right to make their own decisions, and transparency and explainability of the decisions [3, p. 52]. Certainly, if technology and AI will be used to make decisions concerning individuals and groups of people, we will want to ensure this technology addresses ethical concerns, respects humans, and is congruent with the values of the society in which it operates [2, 41].

To this end, the goal, then, is not to stop creating and improving technology, but rather, to create technology with ethical considerations at the forefront of its development–create technology that is cognisant societal responsibilities [1] and operate in a way that an ethical human being would [10]. As Zhong et al. [73] explains, there are three stages to create ethical technology: first, consider how much autonomy to give to its decision-making capacity and what it means to be ethical (source stage); secondly, translate ethical values and implement them (decision stage); lastly, evaluate the technologies impact on people in light of the standard of ethical behaviour (evaluation stage). With these stages in mind, the focus of this work falls under this second 'decision' stage. Given conditions and values have been decided by relevant stakeholders, how might values be implemented into technology to motivate behaviour?

Stakeholder values may be embedded into the behaviour of AI agents through either top-down or bottom-up methodologies [73]. Top-down methodologies are those where the behaviour is formally represented in a logic-based system and reasoned about. While bottom-up methodologies are when desired behaviours are established through data and reasoning is implicit (such as with supervised learning). Top-down approaches have the advantage of being well-defined, which aides in the traceability and explainability in the decision process [73]. Furthermore, top-down approaches may be more reliable for reasoning in comparison to bottom-up AI agents, such as Large Language Models (LLMs), which tend to struggle with irrelevant information and with reasoning that is impacted by errors which cannot currently be combated through up-to-date prompting techniques [60]. Therefore, we consider the top-down approaches to encode stakeholder values into AI agents–systems whose behaviour is automated, either through optimisation or logic programs.

One type of top-down methodology is Value-based Reasoning, a mechanism to justify actions or behaviours using *values* as its central focus. Values, typically being moral values (i.e., power, privacy, autonomy, etc.), play a central role in reasoning to make behaviour that is consistent with the values. The definition of values can be attributed different meanings by various authors, therefore, in §3.2, we investigate the variety of definitions. Values are used in Value-based Reasoning to describe the importance or desires for agents in its decision process. In this way, decisions about actions to take are made by considering how those actions accord with the agent's values. While Value-based Reasoning is a relatively old research topic (cf. Value-based Argumentation [9]), it has become increasingly relevant given the rise of more capable AI, and has demonstrated in many scenarios or applications that include: Artificial Intelligence and Law, where legal cases can be justified with values of the judge or society [12]; privacy and trustworthiness of the hiring process [20]; fairness in water policy [48]; and proportionality in military operations [78].

In this work, we conduct a systematic review of the literature on Value-based Reasoning to gain insights on how researchers create reasoning AI agents and to understand what limitations exist. Specifically, we're interested in how researchers define values and the goals to which determinations are made, and how goals are reached. To aid in scoping this review, we have research questions to be addressed:

RQ1 What is the most accepted definition of 'values' and 'goals' in reasoning systems?
RQ2 How are values represented in AI agents to allow them to reason and make decisions?
RQ3 How are values associated with goals to create a 'motivated agent'?
RQ4 How are values used in dialogues between parties?
RQ5 What are the key problems in the computational analysis of values?

* Corresponding Author. Email: j.p.morgan@swansea.ac.uk
** Corresponding Author. Email: a.z.wyner@Swansea.ac.uk

Table 1. PICO keywords and synonyms used to search digital libraries.

	Keywords	Synonyms
Population	AI	AI Agents, Artificial Intelligence, Machine Learning, Multi-Agent Systems
Intervention	Value-based Reasoning	Value-based argumentation
Comparison	Logic, Argumentation, Reasoning, Philosophy, Value, Computational Model	Norm
Outcome	Representation, Dialogue, Behaviour	Negotiation, Persuasion, Action

An earlier review of Value-based Reasoning was conducted by Guerrero et al. [25] where the authors explored questions such as which theories are used for values and the software/language considered. While, some of our research questions overlap with the work of [25], RQ3-5 examines various facets of Value-based Reasoning that enhance and are complementary to the earlier work.

2 Review Methodology

This review was created following a systematic framework. We first defined the keywords of the topic, which were used to define the search string to query digital libraries for research articles. Finally, relevant articles were selected and information was gathered using a data extraction form.

PICO Keywords The planning stage of the review began with defining a set of PICO (Population, Intervention, Comparison, Outcome) keywords that were related to the research and helped in finding literature to address the research questions [51]. While PICO is typically used in medical fields, it can also be applied to Computer Science [15]. Our keywords and synonyms are presented in Table 1.

Search String From the PICO keywords and their synonyms, we formulated a search string to query various digital libraries:

```
("AI" OR "AI Agents" OR "Artificial Intelligence
    " OR "Machine Learning" OR "Multi-Agent
    Systems")
AND ("Value-based Reasoning" OR "Value-based
    argumentation")
AND ("Argumentation" OR "Computational Model" OR
     "Formal Model" OR "Logic" OR "Philosophy"
    OR "Reasoning" OR "Value" OR "Norm")
AND ("Behaviour" OR "Action" OR "Dialogue" OR "
    Negotiation" OR "Persuasion" OR "
    Representation")
```

Databases The search string was used to query digital libraries:

- ISI Web of Science (http://www.isiknowledge.com);
- Science@Direct (http://www.sciencedirect.com);
- and Scopus(http://www.scopus.com).

In addition, we sourced articles directly from two workshops:

- Value Engineering In AI (VALE) (https://vale2023.iiia.csic.es);
- and Computational Machine Ethics (CME) (https://sites.google.com/view/cme2023/home).

Table 2. The inclusion and exclusion criteria when selecting articles.

Include	Exclude
• Domain of Computer Science • Conference or journal articles • English • Open access • Published in or after the year 2000[1]	• Anonymous authors • Literature reviews with no methodological proposition

Table 3. Questions in the data extraction form.

Question	Data Type
Identified problem (i.e., reason for the work)	String
Solution to the problem	String
Name of logical framework	String
Definition of 'value', 'goal', and 'norm'	String
Hierarchical or preferential treatment of values?	Boolean
How are values linked to goals?	String
Limitations of Framework	String
What scenarios are used as examples of the framework?	String
What is the future works/direction of this work?	String

These workshops were selected as they were directly related to the research in question. As the number of articles in these workshops is smaller relative to digital libraries, we did not need to query them, instead we individually selected relevant articles.

During the query of digital libraries, we posed some criteria which articles must meet to be considered potential articles. Firstly, we used some inclusion and exclusion criteria (Table 2).

Filtering Each potential article was assessed for relevancy based on a preliminary reading of the title and abstract[2]. If the article was deemed relevant, it was selected to become part of our review. After filtering, 57 articles were left for review.

Data Extraction For these 57 articles, a data extraction form with 9 questions was created to address the research questions (Table 3). For each article, an extraction form was filled out and later reviewed with the expanded context of all articles to identify themes.

The overview of the selection process is shown in Fig. 1.

3 Results

In this section, we first perform a quantitive analysis of where the articles have originated from and identify patterns across the corpus of articles. After this, we address the research questions through a thematic analysis of the articles.

3.1 Corpus Statistics

Firstly, we investigate the number of articles published by year. Fig. 2 shows the number of selected articles published was relatively few until 2015. Certainly, by 2023, with the advent of workshops such as VALE, we see a much larger number of articles published. Of course, our review does not cover the entire range of AI ethics, but rather focuses on a smaller or more specific aspect of how to reason with values. Nevertheless, the creation of workshops like VALE and CME represent a shift in thinking about the implications of AI.

[2] A notable theme of articles not selected were related to Preference-based Argumentation Frameworks (PAFs), but were excluded as they do not explicitly consider values as part of their framework. However, if a more expansive review were to take place, we would consider these PAFs to be of relevance even if they do not discuss preferences between values.

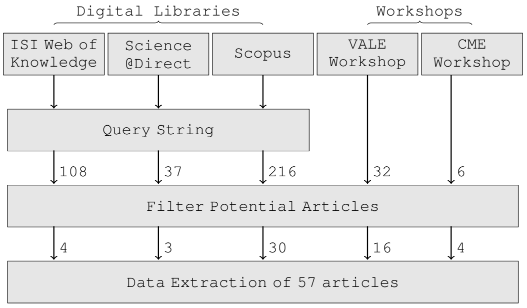


Figure 1. Overview of the article selection process.

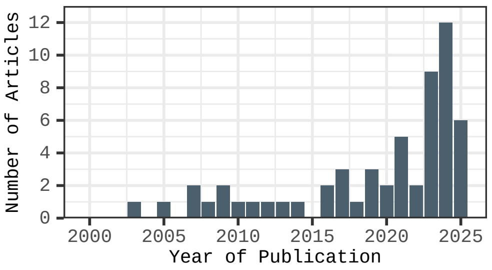


Figure 2. Number of selected articles by publication year.

Fig. 3 (a) shows the top-10 sources with the highest number of articles selected. With 16 articles in our review, VALE is clearly the most represented group. This outcome is predictable in light of the workshop's research objective. The Artificial Intelligence Journal is the second most prevalent, with six articles. Lastly, two conferences–Computational Models of Argument and Legal Knowledge and Information Systems–have three articles each and are ranked third.

The typical number of authors per article is 2 or 3, with the next frequent being 4 authors (Fig. 3 (b)). The minimum number of authors is 1 with the maximum being 8. The overall number of articles per author (regardless of author position) is then examined. Fig. 4 shows the top 20 most prevalent authors. This shows both Trevor Bench-Capon and Tomasz Zurek have the most articles with 8 total. Next Katie Atkinson, Sascha Ossowski, and Adam Wyner with 5.

We analyse the associations of word occurrences in the abstracts of the articles. Words with occurrences of ≥ 5 are selected, leaving 42 salient words in 4 clusters. Fig. 5 shows the association strength between these words. One cluster is dedicated to argumentation (Fig. 5, left), and it's more common associations such as 'attack relation' and 'acceptance', which are more frequently occurring with articles pre-2015 (Fig. 5, right), with 'attack' appearing after 2015. There are numerous references to 'extension' in relation to an extension of Dung's Abstract Argumentation Framework [19]. Later in the timeline, we see the presence of a cluster with words 'social value' and 'goal'. As 'social value' is clustered with 'case' and 'argumentation framework', this social value may reflect values upheld with legal cases where argumentation uses legal cases as examples to demonstrate the frameworks capabilities. The 'goal' keyword is clustered with'plan' which indicates the general usage of the term 'goal'–a goal is realised via a plan of actions. Recent salient words include 'human value', 'value alignment', and 'stakeholder' indicating the increasing complexities which frameworks address. Instead of just 'social values' which typically change less, research is investigating more personal requirements, such as individual or stakeholder values stakeholders. With the rise of 'value alignment', there may be a trend of aligning systems outward behaviour to values instead of directly reasoning with values. The increasing presence of 'human values' may also point existing theories such as Schwartz Theory of Basic Human Values (STBHV) [55] (discussed later in §3.2).

Previous analysis has considered only the abstracts. In this next analysis, we use the full text of the articles. Fig. 6 displays the most common stemmed words appearing in all articles. As to be expected, the most frequent word is 'valuings' with variations being 'value', 'values', 'valuing', etc. Almost all usage of the exact term 'valuings' relate to Belief-Desire-Intention (BDI) where valuings are used to determine preferences among values. Note, there are instances where 'valuings' is used to indicate preferences, leading to conflation of terminology. Words such as 'dialogue', 'conflict', and 'audience' are most often used in articles involving argumentation, in particular value-based argumentation, where dialogue is used to refer to the argumentation between many parties or participants in many moves to propose or deny propositions to reach a conclusion.

As this information is from the scanned texts, words such 'international' may be safely ignored as, after examination, they correspond to footer titles such as 'international conference' etc. However, there are some legitimate uses of 'international', particularly the International Humanitarian Laws considered in the work of Zurek et al. [77].

Finally, for many of the keywords in the visualisation, their presence is unsurprising given their corresponding presence in the search query (§2). However, there are some other clustered keywords which are synonyms to our search query. This indicates the search and filtering method was successful in selecting relevant articles.

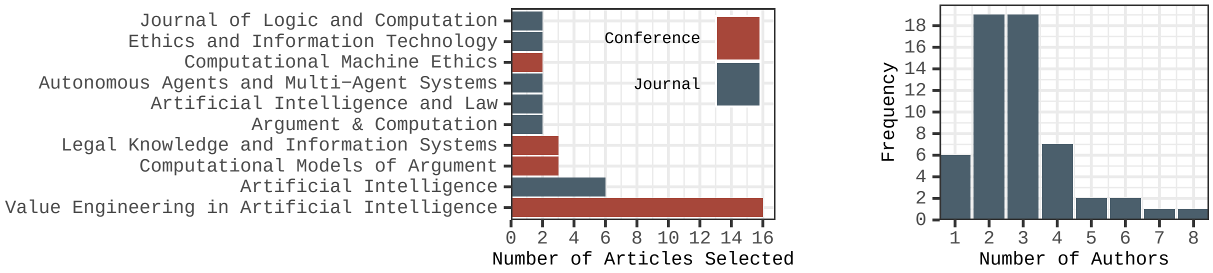
After moving from single words to n-grams, more interesting relationships develop. Fig. 7 presents all two-word n-grams with $\geq$ 40 occurrences. Most prominent is a large cluster of n-grams which typifies the field: abstract-argumentation, argumentation-theories, practical-reasoning, reasoning-chains, decision-process, ethical-decision, etc. We see 'expected-utility' as an expression of consequentialist ethics–a frequent ethical theory considered in the research, in addition to 'deontic-logic' which expresses the formulation of obligatory rules encoded with logic. Meanwhile, there is a cluster related to values with'human-behaviour' and 'human-values' frequently occurring, suggesting research focuses on human values as opposed to any other sort of values. Indeed, we also see values linked with morals, which in turn is linked with community, suggesting a sociological dimension to the values being chosen. Finally, 'values-promoted' is a common for expressing which actions to take–the agent's desire to 'promote' a value by an action.

3.2 Thematic Analysis

In this section, we investigate the emerging themes in the literature. This is organised by research question.

RQ1—What is the most accepted definition of 'values' and 'goals' in reasoning systems? In this research question, we investigate how different authors define the terms 'value' and 'goal' in their works, and the effect these have on the formalisations.

Definition of Values From our investigations, we find: (1) 'value' is often used without an explicit definition. In these cases, we find either the authors use the term 'value' taken as a given, or the article presents their Value-based Reasoning framework in an abstract way such that 'values' can be instantiated later on with a concretization of the framework. In this way, the use of value in these cases is used as placeholders for an idea of preferences with the implicit notion of moral preferences; and (2) the term 'value' is defined explicitly, and it seems, within the context of the selected articles, the



(a) Number of articles per source.

(b) Number of authors per article.

Figure 3. Number of articles published by source (a) and number of authors (b).

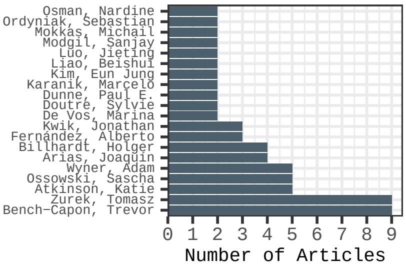


Figure 4. Number of articles per author.

consensus on the definition of 'values' are "abstract principles that guide behaviour" [27, 42]. While this definition may be accepted by many articles, there are differing ideas as to why this may be. Firstly, 'values' may represent abstract standards which define their preferences between states, allowing for the comparison of values, which can be promoted to a certain extent [30, 39, 40, 47, 71, 75, 77, 78].

Other definitions are based on broad motivational goals transcend specific situations or actions and serve as criteria for evaluating the right or wrongness of actions (such as in traditional normative ethics). For these, we see many references to STBHV[3] [22, 23, 26, 29, 49] STBHV is work that defines ten distinct value types evidenced across 16 countries with different cultures [55]. These values are: security, power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, and conformity. The STBHV framework has a specific definition of values, it states: (1) values are beliefs and these values are activated with feelings. While this latter part does not help reasoning systems, it does serve to highlight the motivating factor of people; (2) values refer to desirable goals motivating action towards these goals; (3) values transcend specific situations as the values are general representations of the world desired by value holders; (4) values serve as standards to judge actions in a normative framework; (5) values are ordered by importance relative to one another. This point is shared among the articles in this review where many frameworks (even if they do not refer to STBHV) use ordering relation functions between values; and (6) the relative importance of values to guide action.

While articles relate to Schwartz's definition of values, many articles do not use the whole set of values, which may disrupt the effectiveness of the STBHV framework [25]. As Schwartz [56] points out, these ten values have dynamic relations with one another. For example, benevolence and power conflict, while conformity and security might have compatible relations. It is these tradeoffs between values which lead to behaviours of different people or cultures. However, some recent work address this, for example Oliva-Felipe et al. [49] and Karanik et al. [29], use all 10 STBHV values and use the theory these values relate to one another in complementary or conflicting terms. This allows their methodology to modulate the effect values have on reasoning by the presence of the related values. This way of using STBHV is closer to the idea of the original work.

Definition of Goals The term 'goal' is less defined in literature. However, there is one perception of goals that can be found within the articles. Notably, goals reflects the state of affairs the agent wishes to bring about. In this way, the agent will act accordingly to their values to bring about their preferred or desired state of affairs. However, it is not clear whether these agents really have 'goals' as a perception of a target, or if they make incremental behavioural choices in agreement with their existing values and this just happens to lead to an optimal state of affairs for them. There is a difference here. In the first situation, an agent has a formal representation of a goal, and they use the values at their disposal to reason and bring out the goal and make longer-term planning. In some cases, this may mean taking a sub-optimal short-term action. In the second situation, where the agent doesn't have a formal representation of goals, they are imbued with a set of values reflecting the ideal goal. The difference is whether the agent can reason about goals, or it is incidental to the value system.

However, there are a few articles that give a broader definition to 'goals'. One instance is in Atkinson and Bench-Capon [6] where 4 different types of goals are defined based on the notion of states of propositions, i.e., whether a proposition holds (is true) or not (false): *Achievement Goal*, to make some false true; *Remedy Goal*, to make some true false; *Maintenance Goal*, to keep something currently true, true; and *Avoidance Goal*, to keep something currently false, false. While they have characterised these 4 types of goals, they can still be considered to be related to the state of affairs. Here they are referring to the truthiness of certain states, and the goal is related to how to change the state (or indeed maintain the current state).

Zurek and Mokkas [75] also presents 4 categorisations of goals. *Abstract* and *Unreachable* goals are related to the extent to which values are promoted, with *Unreachable* goals being a subtype of abstract goals where values should be promoted as much as possible. *Material* goals are a situation/state in which an agent satisfy *Abstract* goals. A *Practical* goal is a subtype of *Material* goal which is achievable. Therefore, the categorisations are gradations of *Abstract* goals.

RQ2—How are values represented in AI agents to allow them to reason and make decisions? Throughout the articles, we see two methodologies of creating reasoning systems: (1) through value-alignment meaning values are implicitly encoded into the system through its output behaviour; (2) an explicit representation of values where determinations of inner behaviour and action are being

[3] While there are other competing theories such as Moral Foundations Theory [24], and Values and Identity [54], these do not appear in our articles.

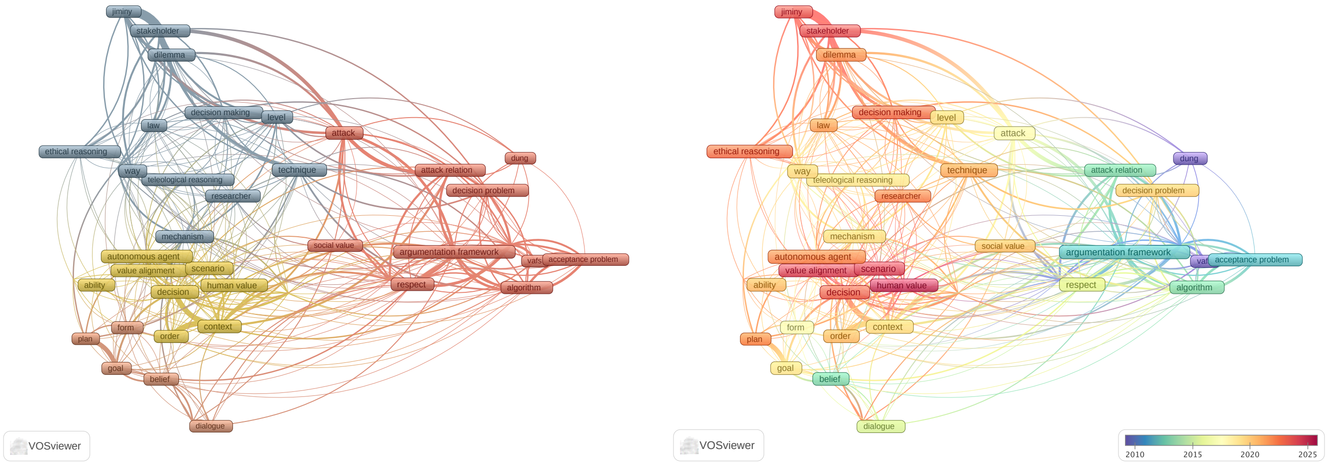


Figure 5. Word association strength between common words in article abstracts, where colour indicates cluster labels (left) and the year of usage (right).

https
graphs
labels
plans
presented
normative
process
proposition
timing
algorithms'
intelligent
assume
bench
extensive
possibly
ways
proposing
letting
considers
important
papers
since
goals
relative
represents
humans
example
actions'
accepts
dialogue
aligns
reasons
systems'
rules'
problems
defining
cases
argument'
differs
need
laws
models
also
international
results
follows
situations
ones
valuings
orders
contexts
autonomous'
ethics
audience
decisive
thus
performs
attacks
agents'
using
two
level
means
theory
states
settings
norm'
logics
formally
practices
givens
prefers
basing
morals
args
capon
works
framework
may
making
sections
allows
defeats
promotions
functions
'social
learns
requiring
approach
computing
particularly
however
definitions
specificity
respects
artificially
weights'
conflicts
provide
discussion

Figure 6. Word cloud of common stemmed words in article text.

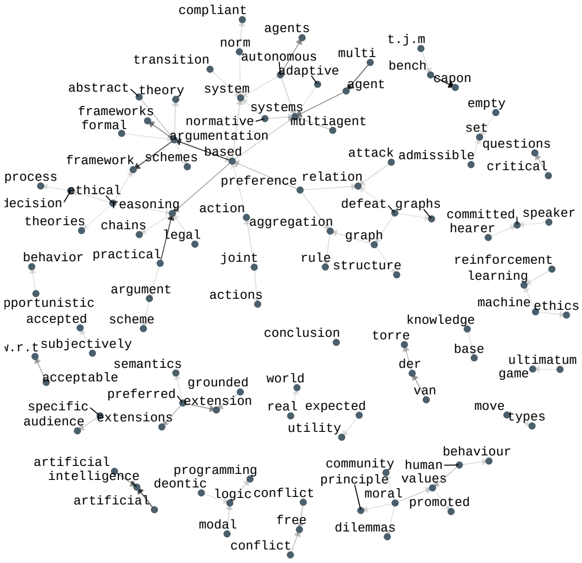


Figure 7. Directed connections between terms in two-word n-grams.

reasoned through states/functions representing the values of interest.

Implicit vs. Explicit Representations With the implicit representation of values, we see systems are made to behave in accordance with values. But these values have no representation within the system in question. In this way, and following the definition of Steels [61], we classify systems are 'value-aligned' where the behaviour (or outcome) of the system reflects the values intended by those who designed the system. Take, for example, the work of Arias et al. [4], where a school placement process is modelled using s(LAW) (based on Answer Set Programming). This application is aimed at improving educational equality, but 'equality' is not explicitly represented within s(LAW), but rather, the propositions such as whether a student has passed a grade threshold or whether they have a disability. The reasoning to achieve educational equality is not based on a particular representation of equality, but on propositions of instrumental value–those that if followed will bring about equality.

Value States When the values are explicitly represented, there are 2 instances in which representations are typically formulated. The first representation of value is that of a state or object representing a value, e.g. [7, 29, 71]. Here we see a set of objects representing a value system, e.g. $V = \{v_1, v_2, ..., v_n\}$ where each v_i represents a particular value, using an STBHV value as an example, power [29]. How these objects get used is different with various methodologies. For example, in Atkinson and Bench-Capon [7], a valuation function defines whether a value gets promoted, demoted, or is unchanged by the transition between two states as in an Action-based Alternating Transition System with Values (AATS+V) [5]. Alternatively, the values can be used to determine an agent's disposition towards a proposition that would, if the metric of disposition passes a threshold, the agent would accept the proposition [70].

Value Functions The second most common formulation–albeit much less common than Value States–is through functions. In this way, there is either one, or a set of functions, representing the values being considered in this work. These functions will typically take a possible future state or action as input and return a real-valued number representing the value of the state w.r.t. to the value (i.e., power) [38, 45]. For example, in Nashed et al. [45], a Markov decision process is used to model the probability of actions leading to state changes in a finite system. This uses a value function to as-

Table 4. Overview of value representations.

Representation	
Implicit	**Explicit**
Value-aligned: [4, 16, 17, 21]	State or Object: [5–8, 11, 13, 14, 22, 23, 28–35, 39, 40, 43, 44, 46, 47, 49, 58, 59, 63–65, 67, 69–72, 77, 79]
	Function: [18, 26, 38, 45, 47, 52, 57, 62, 74–76, 78]
	Numerical: [36, 42, 74, 75]

cribe a 'reward' expected for reaching a particular state. The output of these types of function would then allow the system to reason between many values and choose one that determines the best course of action through a decision procedure. These functions have different names in the literature, including: (1) a *value system function*; (2) a *value semantics function*; or (3) simply a *value function*.

Value Preferences Value states can be accompanied by an ordering relation function to promote value preferences, such that one can form a partial [8, 11, 18, 28, 30, 71] or total [14, 39, 75] ordering between values. Occasionally, the values are represented by a numerical weight or there may be another set in which each value in the set corresponds to a weight for each value. In this case, the value's weight can be used to define the relative strength for the agent [71].

Value Aggregation When many values are considered in the reasoning framework, some mechanism is required to determine the behaviour of the agent. A common theme among logical frameworks is 'value aggregation', where a function is commonly used to compare actions being promoted by various values, ranking them accordingly to a value alignment and selecting the action with the highest alignment [23, 29]. Aggregate functions may also be used to determine the value alignment of a sequence of actions, instead of the alignment of any one single action [26]. This latter version may have the benefit where making a judgment based on the next available actions may inadvertently lead to a suboptimal state than compared with evaluating actions over a series of actions.

Aggregation can be useful for multi-agent systems where conflicting views need to be considered and a determination on how to act made. In this way, argumentation with preference or graph aggregators may be used to choose an outcome for many agents [35], where preference aggregators uses preference relations to determine the orders of values, and a graph aggregator collects many argument graphs into one combined graph to determine the justifiable arguments.

RQ3—How are values associated with goals to create a 'motivated agent'? In earlier works [5], there is a concession that there may not be a need for an explicit or 'ultimate' goal–it may be enough to choose the 'best' action at any given point. In this way, we see there are implicit goals of the agent [6]. However, there has been a development towards a more conscious role of (explicit) goals in the selection process, leading to reason not just about the action at one time, but over a longer period of time [6]. To this end, this research question investigates how, given that values and goals are defined within a reasoning system, these two concepts are connected.

Value Promotion Firstly, if we view goals as being a particular set of state of affairs an agent is motivated to make occur, then, at each step, the agent will decide as to which action to take based on which actions promote or demote its set of values. Take, for example, a crossing a river scenario in Atkinson and Bench-Capon [5], where a farmer wishes to cross the river with seeds, a chicken, and a dog. The (implicit) goal is for the farmer to cross the river with all of these items to make their way home. While, there are shorter term explicit goals, such as making progress (such as getting one of the items to the other side of the river), keeping the chicken and dog happy, and keeping the seeds. At each step, the farmer may decide about what or whom to cross the river with. Using the method of Atkinson and Bench-Capon [5], the farmer makes the decision on which values are promoted or demoted by the decision. In this way, upon making actions that promote values resulting in shorter term goals (such as making progress, or other atomic propositions such as still 'having the chicken', etc.), one makes reach the implicit goal of ultimately crossing the river with all items and animals. This technique of value-promotion as a tool for reaching goals has also been conducted in a system where there are two or more agents by analysing whether cooperation can take place to promote a value [7].

Norms Finally, while values are used to direct the agents' behaviour to a desired state of affairs, norms can be introduced to moderate the sanctioned actions of agents (thereby simplifying representation [11, p. 58]), or help agents to cooperate in a multi-agent system. Norms, like values, have varying definitions across research articles, but a common consensus is that norms mirror social conventions or legal concepts [11, p. 39] in an explicitly ethical system by employing logics and deontic operators. But, while norms can be useful in constraining the actions of agents to sanctioned actions, there may be situations where norms need to be 'broken' as not every possible situation can be anticipated by the norm design [11, p. 51], so mechanisms to allow agents to deny norm-conformance may be necessary. Therefore, to create a motivated agent, explicit norms can be introduced to allow the agent to reason, based on its value system, how might goals be achieved within a normative environment and whether these norms allow for the agent to achieve the goals.

RQ4—How are values used in dialogues between parties? While values can be used to motivate the behaviour of individual agents, we're also interested in how values are used in dialogues or resolve conflicts between many parties/agents.

Audience Firstly, from our survey, we see out of the 57 articles for review, 16 of them discuss the nature of values within dialogues. Most prominently of these works are by Trevor Bench-Capon and his work with Value-based Argumentation Frameworks (VAFs) as an extension on the Abstract Argumentation framework of Dung [19]. In this work, Bench-Capon [13] introduces the concept of audience values into the role of argumentation, motivated by Perelman's New Rhetoric [50] where, for example in a legal dispute, *'each party to a legal dispute 'refers in its argumentation to different values' and the 'judge will allow himself to be guided, in his reasoning, by the spirit of the system, i.e., by the values which the legislative authority seeks to protect and advance'* [50, p. 152]. This brings to focus audiences–they appeal to a different set of values and have different preferences among common values. Any disagreements among audiences is not a logical error, but one of conflicting (often implicit) values and their rankings between values [8]. The role of VAFs is to locate the source of conflicting values among audiences and to attempt to resolve these by finding conflict free arguments (i.e., one without other attacking arguments) which can be satisfiable to audiences.

Value preferences In consequence, in addition to a set of value states (as seen in Research Question 1), VAFs include an additional set of objects representative of the audiences whom each have their preferences to values [7, 8, 46, 58, 59, 72]. The emergence of value preferences are anthropologically motivated in Bench-Capon and

Modgil [11] where the authors note how ‘different norms may evolve in different societies’, and to represent choices as would be acceptable by those societies we ‘need to look at the value order (total or partial) prevailing in that society’ [11, p. 56]. The authors posit that replicating the value preferences of different cultures in the same system may have benefits, though the study of this is left for future work.

Value Promotion From value preferences to argumentation of these value preferences, propositions are put forward by different audiences according to the value they promote [7, 13, 28, 46, 65], which is consistent with Perelman’s original example with the judge’s reasoning of the values established in legislature. However, Kaci and van der Torre [28] state in persuasion dialogues, not all arguments promote values, but can be made based on attacking other values.

Dialogue Types There are examples of various Walton-Krabbe dialogues types [66], each of which have their own initial state and end goal. From the literature, values have been in various dialogue types: *Persuasion*, a conflict on opinion with the goal to resolve the conflict [72]; *Negotiation*, a conflict of interest with a goal to make a beneficial agreement [8, 72]; *Inquiry*, a lack of knowledge with the goal of increasing or adding to the knowledge [72]; *Deliberation*, a need for action with the goal to decide upon an action [8, 72]; *Information-seeking*, the ignorance of knowledge with the goal of understanding the information [72]; and finally *Eristic*, an antagonism between parties with the goal to accommodate these parties [72]. Other dialogue types do not appear in the considered articles.

Limitations Lastly, while values may be used for resolving conflicts in dialogues via argumentation frameworks (such as VAFs), the significant issue is the selection of acceptable arguments satisfiable for many audiences. As described in Kim et al. [31], for arguments of value-width 2 and attack-width 1, finding a subjective acceptance is an NP-hard problem, thereby limiting its usage in realistic scenarios.

RQ5—What are the key problems in the computational analysis of values? Lastly, we evaluate the key issues with the formalisations that might point to the future direction of the field.

Complexity In among the selected articles, the most common limitation is complexity, i.e., the complexity required for a reasoning system to operate in a real-world environment. Take a typical consequentialist-style system, for example, where reasoning is made by considering the consequence of such a behaviour. In these systems, evaluating and comparing the preferences between various state transitions for many value functions may not be realizable within a real-time operation. However, while this limitation has been identified, it may not be necessary, or indeed even the intention of the author’s work to create such a framework that is capable of operating in a real-world autonomous system. It may be, for example, the authors wish to explore the mechanisms behind reasoning and concretise the logic formulation, allowing them to explore and analyse how a system may act in different situations or consider the propositions involved in the option selection process. In this case, they would be forgiven if the system was not tractable within a real-time environment, as it was not ultimately a concern for the work.

Generalisability The second-most-common problem is generalisability. For articles where formulations are designed for a specific application, we would not expect for the work to transcend the application in question. In other cases, the design of the framework is built in abstract, with examples showing how the framework may be instantiated with a set of values or goals to achieve. Despite this, theoretical claims about the plausibility of the framework’s usability, or flexibility in the definitions [70], is left to be tested in future work, as there may be a lack of more realistic scenarios or data [27]. Though there are common scenarios for testing Value-based Reasoning, future work might consider building realistic dilemmas and data, allowing for comparative analysis between proposed methods.

Connection between Values and Goals In some cases, the goals of an agent have been made implicit, and by the design of the agent’s operation it will be made to achieve the goal [5]. However, having explicit goals may have the benefit of allowing the agent to (flexibly) reason as to how to best achieve these goals in relation to its value system [6]. Furthermore, future work should consider what is meant by goals. Indeed, there has recent work in this area where goals have been categorised depending on how values are promoted or achieved [75], but more could be done in this area to solidify the relationship between values and goals. One may look at the literature concerning long-term planning and how this is formalised to better distinguish between goals in the short-term versus the long term.

Concept of Preferences Preferences so far in the considered literature have been used as a way of resolving conflicts between different actions that promote different values by evaluating the relative strength of values between one another, ultimately leading to the ‘best’ action to take. However, it is not fully clear in what way preferences exist in relation to values. For example, if ‘I prefer apples to having oranges‘, here there is a preference of in favour of the value of having apples. But, one may question why this is in the first instance. Perhaps it is due to the taste or texture, thus the interrogation of the preference has led to the discovery of more important values such ‘taste’. So in the context of making a choice between apples and oranges, it is not the preference between the two shallow values of ‘having apples‘ and ‘having oranges’, but rather to which decision leads to the promotion of the more important ‘taste’ value. Despite this, there would still need to be a computational representation of importance between values to model ‘taste’ versus ‘having apples’.

In addition, it is not always clear if preferences should remain consistent and be predefined. The analysis of preferences and values may be a benefit from integrating contextual information, such as consequences of taking action at the current state [68]. Future work should consider how preferences and values are adapted to different states.

Finally, it is argued personal preferences shouldn’t be the only things that matter when deciding upon actions, but one should also consider other properties such as compliance to norms that may override any subjective preferences [37] – to create good social policy, personal preferences must not override social interests [53].

4 Conclusion

Value-based reasoning, as an abstract methodology for incorporating moral values into the decision process of automated or AI agents, is becoming more prevalent in research as the capabilities of AI technologies increases. In this work, we have conducted a review of the literature on Value-based Reasoning. This review has used a systematic methodology to find and extract information from the articles, to which we apply a thematic and corpus analysis. From this review, we identify patterns among research such as the representation of values and how these values are used in deciding the behaviour of agents. Finally, while Value-based Reasoning can be an applicable method for creating trustworthy AI by incorporating and respecting the values of stakeholders, we foresee more work to be done on formalising the interaction between values, goals, and preferences. Future work may want to consider these three aspects.

Acknowledgements

This work was funded by Swansea University. For the purpose of Open Access, the author has applied a CC BY licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

[1] IEEE Standard Model Process for Addressing Ethical Concerns during System Design. *IEEE Std 7000-2021*, pages 1–82, Sept. 2021. doi: 10.1109/IEEESTD.2021.9536679.

[2] N. Ajmeri, H. Guo, P. K. Murukannaiah, and M. P. Singh. Designing ethical personal agents. *IEEE Internet Computing*, 22(2):16–22, 2018.

[3] M. Anderson and S. L. Anderson, editors. *Machine Ethics*. Cambridge University Press, Cambridge, 2011. ISBN 978-0-521-11235-2. doi: 10.1017/CBO9780511978036.

[4] J. Arias, M. Moreno-Rebato, J. A. Rodriguez-García, and S. Ossowski. Value Awareness and Process Automation: A Reflection Through School Place Allocation Models. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 261–269, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_16.

[5] K. Atkinson and T. Bench-Capon. Practical Reasoning as Presumptive Argumentation Using Action Based Alternating Transition Systems. *Artificial Intelligence*, 171(10):855–874, 2007. ISSN 0004-3702. doi: 10.1016/j.artint.2007.04.009.

[6] K. Atkinson and T. Bench-Capon. States, Goals and Values: Revisiting Practical Reasoning. *Argument & Computation*, 7(2-3):135–154, 2016. ISSN 1946-2166. doi: 10.3233/AAC-160011.

[7] K. Atkinson and T. Bench-Capon. Taking Account of the Actions of Others in Value-Based Reasoning. *Artificial Intelligence*, 254:1–20, 2018. ISSN 0004-3702. doi: 10.1016/j.artint.2017.09.002.

[8] K. Atkinson, T. Bench-Capon, and P. McBurney. A Dialogue Game Protocol for Multi-Agent Argument Over Proposals for Action. *Autonomous Agents and Multi-Agent Systems*, 11(2):153–171, 2005. ISSN 1573-7454. doi: 10.1007/s10458-005-1166-x.

[9] T. Bench-Capon. Value Based Argumentation Frameworks. *arXiv preprint cs/0207059*, July 2002.

[10] T. Bench-Capon. Ethical approaches and autonomous systems. *Artificial Intelligence*, 281:103239, 2020. ISSN 0004-3702. doi: 10.1016/j.artint.2020.103239.

[11] T. Bench-Capon and S. Modgil. Norms and Value Based Reasoning: Justifying Compliance and Violation. *Artificial Intelligence and Law*, 25(1):29–64, 2017. ISSN 1572-8382. doi: 10.1007/s10506-017-9194-9.

[12] T. Bench-Capon and H. Prakken. A case study of hypothetical and value-based reasoning in US Supreme-Court cases. In *Legal Knowledge and Information Systems*, pages 11–20. IOS Press, 2009.

[13] T. J. Bench-Capon. Persuasion in Practical Argument Using Value-Based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003. ISSN 0955-792X. doi: 10.1093/logcom/13.3.429.

[14] T. J. Bench-Capon, S. Doutre, and P. E. Dunne. Audiences in Argumentation Frameworks. *Artificial Intelligence*, 171(1):42–71, 2007. ISSN 0004-3702. doi: 10.1016/j.artint.2006.10.013.

[15] A. Carrera-Rivera, W. Ochoa, F. Larrinaga, and G. Lasa. How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9:101895, 2022. ISSN 22150161. doi: 10.1016/j.mex.2022.101895.

[16] L. R. Cima, D. De Jonge, and N. Osman. Towards the Incorporation of Social Values in Automated Negotiation Strategies. In *Value Engineering in Artificial Intelligence. VALE 2024*, volume 15356 of *Lecture Notes in Computer Science*, pages 193–207, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_12.

[17] G. Dalmasso, L. Marcos-Vidal, and C. Pretus. Modelling Moral Decision-Making in a Contractualist Artificial Agent. In *Value Engineering in Artificial Intelligence. VALE 2024*, volume 15356 of *Lecture Notes in Computer Science*, pages 155–175, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_10.

[18] L. A. Dennis and C. P. del Olmo. A Defeasible Logic Implementation of Ethical Reasoning. In *First International Workshop on Computational Machine Ethics (CME-2021)*, 2021.

[19] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995. ISSN 0004-3702. doi: 10.1016/0004-3702(94)00041-X.

[20] C. Fernández-Martínez and A. Fernández. Value-Based Reasoning Scenario in Employee Hiring and Onboarding Using Answer Set Programming. In N. Osman and L. Steels, editors, *Value Engineering in Artificial Intelligence*, volume 14520, pages 251–260. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-58204-2 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_15.

[21] C. Fernández-Martínez and A. Fernández. Value-Based Reasoning Scenario in Employee Hiring and Onboarding Using Answer Set Programming. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 251–260, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_15.

[22] M. Garcia-Bohigues, C. Cordova, J. Taverner, J. Palanca, E. Del Val, and E. Argente. Towards a Distributed Platform for Normative Reasoning and Value Alignment in Multi-Agent Systems. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 237–250, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_14.

[23] S. García-Rodríguez, M. Karanik, and A. Pina-Zapata. Value Promotion Scheme Elicitation Using Natural Language Processing: A Model for Value-Based Agent Architecture. In *Value Engineering in Artificial Intelligence. VALE 2024*, volume 15356 of *Lecture Notes in Computer Science*, pages 104–120, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_7.

[24] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

[25] E. Guerrero, S.-T. Tzeng, C. Pastrav, and F. Dignum. Value-Based Decision-Making in Software Agents: A Systematic Literature Review. In N. Osman and L. Steels, editors, *Value Engineering in Artificial Intelligence*, volume 15356, pages 137–154. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-85462-0 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_9.

[26] A. Holgado-Sánchez, J. Arias, H. Billhardt, and S. Ossowski. Algorithms for Learning Value-Aligned Policies Considering Admissibility Relaxation. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 145–164, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_9.

[27] A. Holgado-Sánchez, J. Bajo, H. Billhardt, S. Ossowski, and J. Arias. Value Learning for Value-Aligned Route Choice Modeling Via Inverse Reinforcement Learning. In *Value Engineering in Artificial Intelligence. VALE 2024*, volume 15356 of *Lecture Notes in Computer Science*, pages 40–60, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_3.

[28] S. Kaci and L. van der Torre. Preference-Based Argumentation: Arguments Supporting Multiple Values. *International Journal of Approximate Reasoning*, 48(4):730–751, 2008. ISSN 0888-613X. doi: 10.1016/j.ijar.2007.07.005.

[29] M. Karanik, H. Billhardt, A. Fernández, and S. Ossowski. Exploiting Value System Structure for Value-Aligned Decision-Making. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 180–196, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_11.

[30] E. J. Kim and S. Ordyniak. Valued-Based Argumentation for Tree-Like Value Graphs. In *Computational Models of Argument*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 378–389, 2012. doi: 10.3233/978-1-61499-111-3-378.

[31] E. J. Kim, S. Ordyniak, and S. Szeider. Algorithms and Complexity Results for Persuasive Argumentation. *Artificial Intelligence*, 175(9): 1722–1736, 2011. ISSN 0004-3702. doi: 10.1016/j.artint.2011.03.001.

[32] S. Kolker, L. Dennis, R. F. Pereira, and M. Xu. Uncertain Machine Ethical Decisions Using Hypothetical Retrospection. In *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, pages 161–181, Cham, 2023. Springer Nature Switzerland.

[33] C. Leturc and G. Bonnet. Using N-Ary Multi-Modal Logics in Argumentation Frameworks to Reason About Ethics. *AI Communications*, 37(3):323–355, 2024. ISSN 0921-7126. doi: 10.3233/AIC-220301.

[34] B. Liao, P. Pardo, M. Slavkovik, and L. van der Torre. The Jiminy Advisor: Moral Agreements Among Stakeholders Based on Norms and Argumentation. *Journal of Artificial Intelligence Research*, 77:737–792, 2023. ISSN 1076-9757. doi: 10.1613/jair.1.14368.

[35] G. Lisowski, S. Doutre, and U. Grandi. Aggregation in Value-Based Argumentation Frameworks. In *Electronic Proceedings in Theoretical Computer Science*, volume 297, pages 313–331, 2019. ISBN 2075-2180. doi: 10.4204/eptcs.297.20.

[36] A. López-García. A Proposal for Selecting the Most Value-Aligned Preferences in Decision-Making Using Agreement Solutions. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, volume 1: EAA, pages 461–470. SciTePress, 2024. ISBN 978-989-758-680-4. doi: 10.5220/0012586300003636.

[37] A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable. Preferences and ethical principles in decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 222–222, 2018.

[38] M. Lujak, A. Fernández, H. Billhardt, S. Ossowski, J. Arias, and A. López Sánchez. On Value-Aligned Cooperative Multi-Agent Task Allocation. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 197–216, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_12.

[39] J. Luo, J.-J. Meyer, and M. Knobbout. A Formal Framework for Reasoning About Opportunistic Propensity in Multi-Agent Systems. *Autonomous Agents and Multi-Agent Systems*, 33(4):457–479, 2019. ISSN 1573-7454. doi: 10.1007/s10458-019-09413-1.

[40] J. Luo, B. Liao, and D. Gabbay. Value-Based Practical Reasoning: Modal Logic + Argumentation. In *Computational Models of Argument*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, pages 248–259, 2022. doi: 10.3233/FAIA220157.

[41] D. K. McGraw. Ethical Responsibility in the Design of Artificial Intelligence (AI) Systems. *International Journal on Responsibility*, 7(1), Nov. 2024. ISSN 2576-0955. doi: 10.62365/2576-0955.1114.

[42] R. Mercuur, V. Dignum, and C. Jonker. The Value of Values and Norms in Social Simulation. *Journal of Artificial Societies and Social Simulation*, 22(1), 2019. ISSN 1460-7425. doi: 10.18564/jasss.3929.

[43] S. Modgil. Reasoning About Preferences in Argumentation Frameworks. *Artificial Intelligence*, 173(9):901–934, 2009. ISSN 0004-3702. doi: 10.1016/j.artint.2009.02.001.

[44] N. Montes, N. Osman, and C. Sierra. Perspective-Dependent Value Alignment of Norms. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 46–63, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_4.

[45] S. Nashed, J. Svegliato, and S. Zilberstein. Ethically Compliant Planning Within Moral Communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 188–198. Association for Computing Machinery, 2021. doi: 10.1145/3461702.3462522.

[46] S. Nofal, K. Atkinson, and P. E. Dunne. Algorithms for Decision Problems in Argument Systems Under Preferred Semantics. *Artificial Intelligence*, 207:23–51, 2014. ISSN 0004-3702. doi: 10.1016/j.artint.2013.11.001.

[47] P. Noriega and E. Plaza. On Autonomy, Governance, and Values: An AGV Approach to Value Engineering. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 165–179, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_10.

[48] L. Oliva-Felipe, I. Lobo, J. McKinlay, F. Dignum, M. De Vos, U. Cortés, and A. Cortés. Context Matters: Contextual Value-Based Deliberation in Water Consumption Scenarios. In N. Osman and L. Steels, editors, *Value Engineering in Artificial Intelligence*, volume 15356, pages 208–222. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-85462-0 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_13.

[49] L. Oliva-Felipe, I. Lobo, J. McKinlay, F. Dignum, M. De Vos, U. Cortés, and A. Cortés. Context Matters: Contextual Value-Based Deliberation in Water Consumption Scenarios. In *Value Engineering in Artificial Intelligence. VALE 2024*, volume 15356 of *Lecture Notes in Computer Science*, pages 208–222, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_13.

[50] C. Perelman. *The New Rhetoric and the Humanities*. Springer Netherlands, Dordrecht, 1979. ISBN 978-90-277-1019-2 978-94-009-9482-9. doi: 10.1007/978-94-009-9482-9.

[51] K. Petersen, S. Vakkalanka, and L. Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, 2015. ISSN 0950-5849. doi: 10.1016/j.infsof.2015.03.007.

[52] M. Riad, S. Ghanadbashi, and F. Golpayegani. Run-Time Norms Synthesis in Dynamic Environments with Changing Objectives. In *Artificial Intelligence and Cognitive Science*, volume 1662 CCIS, pages 462–474. Springer, Cham, 2023. ISBN 978-3-031-26438-2. doi: 10.1007/978-3-031-26438-2_36.

[53] M. Sagoff. Values and preferences. *Ethics*, 96, 01 1986. doi: 10.1086/292748.

[54] M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, D. PINHO, A. H. VINAGREIRO, E. Vecchione, L. Scheunemann, et al. Values and identities-a policymaker's guide. 2021.

[55] S. H. Schwartz. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Elsevier, 1992. ISBN 978-0-12-015225-4. doi: 10.1016/S0065-2601(08)60281-6.

[56] S. H. Schwartz. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1), Dec. 2012. ISSN 2307-0919. doi: 10.9707/2307-0919.1116.

[57] M. Serramia, M. Rodriguez-Soto, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, F. Bistaffa, P. Boddington, M. Wooldridge, and C. Ansotegui. Encoding Ethics to Compute Value-Aligned Norms. *Minds & Machines*, 33(4):761–790, 2023. ISSN 1572-8641. doi: 10.1007/s11023-023-09649-7.

[58] Z. Shams, M. De Vos, N. Oren, and J. Padget. Argumentation-Based Reasoning About Plans, Maintenance Goals, and Norms. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 14(3):1–39, 2020. ISSN 1556-4665. doi: 10.1145/3364220.

[59] M. Snaith. An Argument-Based Framework for Selecting Dialogue Move Types and Content. In *Computational Models of Argument*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 355–362, 2020. doi: 10.3233/FAIA200519.

[60] S. H. Song and W. Tavanapong. How much do prompting methods help llms on quantitative reasoning with irrelevant information? In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2128–2137, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679840. URL https://doi.org/10.1145/3627673.3679840.

[61] L. Steels. Values, Norms and AI. In N. Osman and L. Steels, editors, *Value Engineering in Artificial Intelligence*, pages 1–7, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_1.

[62] J. Svegliato, S. B. Nashed, and S. Zilberstein. Ethically Compliant Sequential Decision Making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 of *AAAI Technical Track on Philosophy and Ethics of AI*, pages 11657–11665, 2021. doi: 10.1609/aaai.v35i13.17386.

[63] Y. S. Taheri, G. Bourgne, and J.-G. Ganascia. Modelling Integration of Responsible AI Values for Ethical Decision Making. In *KR 2023 Workshop on Computational Machine Ethics*, 2023.

[64] B. Verheij. Formalizing Value-Guided Argumentation for Ethical Systems Design. *Artificial Intelligence and Law*, 24(4):387–407, 2016. ISSN 1572-8382. doi: 10.1007/s10506-016-9189-y.

[65] D. Walton. A Dialogue Model of Belief. *Argument & Computation*, 1 (1):23–46, 2010. doi: 10.1080/19462160903494576.

[66] D. Walton and E. C. Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, 1995.

[67] R. Wannous and C. Trojahn. Explaining Argumentation Over Alignment Agreements. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 80–85, 2013. doi: 10.1109/WI-IAT.2013.94.

[68] J. Woodgate. Ethical principles for reasoning about value preferences. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 972–974, 2023.

[69] A. Wyner and T. Bench-Capon. Modelling Judicial Context in Argumentation Frameworks. *Journal of Logic and Computation*, 19(6):941–968, 2009. ISSN 0955-792X. doi: 10.1093/logcom/exp009.

[70] A. Wyner and T. Zurek. On Legal Teleological Reasoning. In *Legal Knowledge and Information Systems*, volume 379 of *Frontiers in Artificial Intelligence and Applications*, pages 83–88, 2023. doi: 10.3233/FAIA230948.

[71] A. Wyner and T. Zurek. Towards a Formalisation of Motivated Reasoning and the Roots of Conflict. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 28–45, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_3.

[72] A. Wyner and T. Zurek. Satisfaction in Negotiation by Structured Values and Propositions. In *Value Engineering in Artificial Intelligence. VALE 2024*, volume 15356 of *Lecture Notes in Computer Science*, pages 176–192, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-85463-7. doi: 10.1007/978-3-031-85463-7_11.

[73] T. Zhong, Y. Song, R. Limarga, and M. Pagnucco. Computational Ma-

chine Ethics: A Survey. *Journal of Artificial Intelligence Research*, 82: 1581–1628, Mar. 2025. ISSN 1076-9757. doi: 10.1613/jair.1.16836.

[74] T. Zurek and M. Mokkas. Modeling Value-Based Reasoning for Autonomous Agents. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 259–262. IEEE, 2017. doi: 10.15439/2017F282.

[75] T. Zurek and M. Mokkas. Value-Based Reasoning in Autonomous Agents. *International Journal of Computational Intelligence Systems*, 14(1):896–921, 2021. ISSN 1875-6883. doi: 10.2991/IJCIS.D.210203.001.

[76] T. Zurek, M. Mohajeriparizi, J. Kwik, and T. Van Engers. Can a Military Autonomous Device Follow International Humanitarian Law? In *Legal Knowledge and Information Systems*, volume 362 of *Frontiers in Artificial Intelligence and Applications*, pages 273–278, 2022. doi: 10.3233/FAIA220479.

[77] T. Zurek, J. Kwik, and T. van Engers. Model of a Military Autonomous Device Following International Humanitarian Law. *Ethics and Information Technology*, 25(1):15, 2023. doi: 10.1007/s10676-023-09682-1.

[78] T. Zurek, J. Kwik, and T. Van Engers. Values, Proportionality, and Uncertainty in Military Autonomous Devices. In *Value Engineering in Artificial Intelligence. VALE 2023*, volume 14520 of *Lecture Notes in Computer Science*, pages 219–236, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58202-8. doi: 10.1007/978-3-031-58202-8_13.

[79] T. Zurek, A. Wyner, and T. Bench-Capon. Values and Factor Ascription Arguments. In *Legal Knowledge and Information Systems*, volume 395 of *Frontiers in Artificial Intelligence and Applications*, pages 239–248, 2024. doi: 10.3233/FAIA241249.