An Ontology for Value Awareness Engineering

Andrés Holgado-Sánchez[a], Holger Billhardt[b], Sascha Ossowski[c] and Alberto Fernández[d]
CETINIA, Universidad Rey Juan Carlos, Madrid, Spain

Keywords: Value Awareness Engineering, Value-Alignment, Ontology Engineering.

Abstract: The field of *value awareness engineering* claims that intelligent software agents should be endowed with a set of capabilities related to human values, enabling them to identify value-aligned outcomes and, ultimately, to choose their behaviour in value-aware manner. In this work we develop an ontology that links many of the models and concepts that have been proposed in relation to computational value awareness, so as to be able to formalize in a common language the various heterogeneous research proposals in the field. Specifically, we illustrate its capability for describing multi-agent systems from the value-awareness engineering perspective through several case studies grounded in concrete approaches from literature. The ontology, implemented in OWL and extended with SWRL rules, is evaluated following scenarios of the NeOn Methodology and is interconnected with relevant ontologies in the Semantic Web.

1 INTRODUCTION

AI systems that explicitly represent and reason with human values have recently been studied in the new research field of *value-awareness engineering* (VAE) (Sierra et al., 2021; Montes et al., 2023). The VAE field covers various approaches to formally design *value-aware systems*, i.e. systems involving cognitive agents that are provided with mechanisms to behave according to values and being able to reason with them; assessing the feasibility of executing different behaviours, selecting reasonable norms or following different goals in terms of their *value-alignment* (Russell, 2022; Rodriguez-Soto et al., 2022; Balakrishnan et al., 2019); caring about specific value relationships (i.e. formalizing and understanding *value systems* (Lera-Leri et al., 2022; Serramia et al., 2018)); and taking into account (or learning) their context and agent-dependent nature (i.e. that different agents may hold different preferences in different situations (Montes and Sierra, 2022; Osman and d'Inverno, 2023; Sierra et al., 2021; Soares, 2018)).

Despite the undeniably fruitful research made so far, the diversity of proposals leads to an increasing heterogeneity in the nomenclature in the field, mostly due to application-biased interpretations of values from different computational and social science theories (e.g. (Montes and Sierra, 2022; Osman and d'Inverno, 2023) with (Schwartz, 1992), (Lera-Leri et al., 2022) with (Chisholm, 1963) or (Rodriguez-Soto et al., 2022) with (Arnold et al., 2017)). Though first efforts have been put forward towards the "formalization of the moral and social values as *abstract objects with social capital*" (De Giorgis et al., 2022), there is still a lack of a *common-language* around even basic notions in value-aware AI. For instance, there are different notions of *norms* (Serramia et al., 2020; Serramia et al., 2018; Rodriguez-Soto et al., 2022); different ways to *ground values* (i.e. the specific ways of evaluating values under specific problems) and value systems (Serramia et al., 2018; Osman and d'Inverno, 2023); and diverse definitions of value-alignment of AI behaviours, referring to either *actions* and *strategies* (deontological view, (Lera-Leri et al., 2022; Rodriguez-Soto et al., 2022)) or on *states* and/or sequences of decisions (consequentialist view, (Montes and Sierra, 2022; Holgado-Sánchez et al., 2023)).

Recent proposals from the *value awareness engineering* field were discussed within the VALE workshop celebrated at ECAI 2023 (Steels, 2023)[1]. In particular, interesting discussions were spawned regarding the acceptation of high level concepts related to the field.

[a] https://orcid.org/0000-0001-8853-1022
[b] https://orcid.org/0000-0001-8298-4178
[c] https://orcid.org/0000-0003-2483-9508
[d] https://orcid.org/0000-0001-8298-4178

[1] VALE-2023 pre-proceedings, Osman et al. (eds.): https://vale2023.iiia.csic.es/accepted-papers

Holgado-Sánchez, A., Billhardt, H., Ossowski, S. and Fernández, A.
An Ontology for Value Awareness Engineering.
DOI: 10.5220/0012595500003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 1421-1428
ISBN: 978-989-758-680-4; ISSN: 2184-433X

Inspired by that effort to advance towards a terminological consensus, in this paper we propose the 'VAE ontology"[2] (based on OWL[3]). The goal of this ontology is to represent with a common vocabulary different concepts and notions that the VAE community has proposed so far regarding the design of value-aware agent-based systems, their relations and the underlying formalisms. It aims at reducing the gap between the theoretical experiments (and the theory itself) and implemented prototypes, providing interoperability with regard to different theories. To validate our proposed ontology, we analyze in detail three proposals from the literature, regarding consequentialist *value-aligned norm selection* (Montes and Sierra, 2022), value representation using taxonomies (Osman and d'Inverno, 2023), and deontological *value-aligned norm selection* (Serramia et al., 2018), and show how these approaches can be modelled and represented within the VAE ontology.

The paper is structured as follows. Section 2 compiles related work regarding computational value-awareness and ontologies. In section 3, we describe in detail the different parts of the VAE ontology. Section 4 illustrates how different research proposals from the literature can be modelled within the VAE ontology. Finally, section 5 discusses some of the lessons learnt, while section 6 concludes the paper and outlines avenues for future research.

2 RELATED WORK

2.1 Value-Awareness Engineering

According to (Poole and Mackworth, 2010), in a decision problem, for intelligent agents to know which action to take, they should understand the effects of each action and the preferences they have over their effects. Human values should certainly have an impact over these preferences, but assessing that impact has turned out to be notoriously difficult.

Still, approaches to incorporate specific values into the reasoning and decision-making schemes of intelligent software agents have been developed. These proposals date back from practical reasoning, pioneers using the notion of *value systems* in argumentation systems (Bench-Capon et al., 2012), defined via value preferences over states and or actions. This was then used later in various original problems such as finding value-aligned normative systems (Serramia et al., 2020; Montes and Sierra, 2022; Montes and Sierra, 2021); analyzing the value-alignment of outcomes (value-aware decision making) (Sierra et al., 2021; Rodriguez-Soto et al., 2022), value aggregation (aggregating agents preferences into ranked values) (Lera-Leri et al., 2022); and *value learning* (Soares, 2018), i.e. learning representations of values, by classifying (potential) outcomes.

However, most differ in their understanding of values. Some authors in the VAE community advocate for a *consequentialist view* (Montes and Sierra, 2022), mostly inspired by Schwartz's theory of Basic Human Values (Schwartz, 1992). The key assumption is that "values serve as standards, refer to desirable goals and transcend specific actions" (Sierra et al., 2021), which a similar stance than that of (Poole and Mackworth, 2010).

Other authors prefer a deontological approach, stating that the actions or norms have an intrinsic meaning related to values (Lera-Leri et al., 2022; Serramia et al., 2020) and not the results of their application. Others are skeptic defining such intrinsic relationships between outcomes (or even goals) with values (Soares, 2018; Osman and d'Inverno, 2023).

2.2 Ontologies for Value-Aware Systems

The justification of using an ontology to represent value reasoning theories is sustented by (Soares, 2018), where he presents the problem of *ontology identification* as essential for the value learning problem. The approach relies on learning an ontology that reflects the knowledge that agents need to classify outcomes according to values in changing contexts.

However, work on ontologies modelling or supporting value-awareness in AI is scarce. A notable exception is the ValueNet ontology network (De Giorgis et al., 2022), "a modular ontology representing and operationalising moral and social values" corresponding to different value theories, namely "Basic Human Values"[4] (Schwartz, 1992) and "Moral Foundations Theory"[5] (Graham et al., 2013). The goal of that ontology was finding moral content in human discourse.

Despite the lack of ontologies considering human values, there is a variety of ontologies formalizing relevant notions in the VAE literature, namely, the notion of social *norms* that regulate agent behaviour (ideally being aligned with our values); *agents* that operate in line with them; and *outcomes* that occur or are provoked in the system. For instance, the OWL-based ontology NIC (Gangemi, 2008) modelled interactions between *agents*, *plans* and *norms*. In a similar line, (Fornara and Colombetti, 2010) developed an

[2]VAE ontology IRI: https://w3id.org/def/vaeontology
[3]OWL https://www.w3.org/TR/owl-guide/
[4]https://w3id.org/spice/SON/SchwartzValues
[5]https://w3id.org/spice/SON/HaidtValues

OWL application-independent ontology with SWRL rules (Grosof et al., 2003) conveying temporal propositions, *events*, agents, roles, norms and social commitments. Another example of an ontology for normative specification is the ODRL Information Model 2.2 (Ianella and Villata, 2018), a W3C recommended ontology for representing statements about the rights of usage of content and services.

3 THE VAE ONTOLOGY

The goal for the VAE ontology is providing a common representation for VAE theories usable in agent-based value-aware and normative systems. To develop it, we followed the *NeOn methodology* for *Ontology Engineering* (Suárez-Figueroa et al., 2015), specially designed for building ontology networks. It comprises a series of activities[6] designed for different scenarios. Here we summarize the key activities performed.

First, we performed the specification of competency questions (CQs) (i.e. the functional requirements), summarized in the following list.

CQ1. What is the definition of a *value-related concept*[7] depending on the context and, if it is part of a theory (e.g. BHV), what is its classification?

CQ2. How do *norms* affect *agents*?

CQ3. What type of *outcomes* (events) exist and what agents *participate* in them?

CQ4. What *properties* of an outcome or norm are related with a given value and in which context?

CQ5. What statements can an agent propose about *alignment* of outcomes/norms with values, by looking at what *properties*?

CQ6. What agents are stated to be value-aware, according to properties of their behaviour?

CQ7. What arguments[8] an agent proposes to support its value statements about the world?

Then, we investigated reusable ontological resources. highlighting: first, parts of the DOLCE+DnS Ultralite ontology, a general-purpose and lightweight version of DOLCE (Gangemi et al., 2003) (from the authors of NIC), where we root our notions

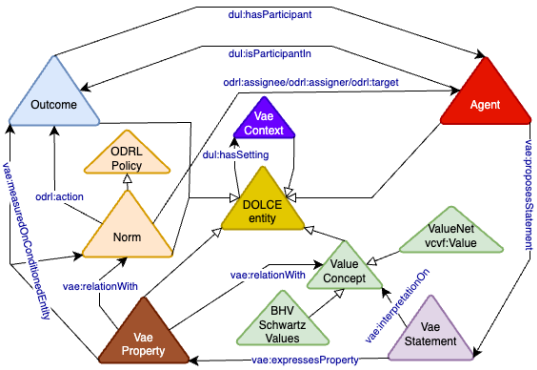

Figure 1: Schematic conceptual diagram of the VAE ontology with the most relevant relations between high level concepts. Colors are used to identify groups of similar notions in the ontology.

for norm (`dul:Norm`), agent (`dul:Agent`), outcome (`dul:Event`), statement (`dul:Description`) and context (`dul:Situation`); second, classes from the ODRL ontology for norm specification; and third, all values from the BHV ontology from ValueNet, and their of value (`vcvf:Value`).

The next activities were the *conceptualization* and *formalization* of the ontology. The design of the VAE ontology was conceived as a *core module* to which other ontologies representing different specific proposals are aligned to. To guarantee essential interoperability in the Semantic Web, we opted for OWL as the implementation language, aided by SWRL rules.

Finally, the implementation was subject to an *evaluation* process, that consisted of checking the complete representation of the CQs, checking design pitfalls with OOPS! (Poveda-Villalón et al., 2014) and assessing the correctness of the ontology using the Pellet (Sirin et al., 2007) reasoner.

To illustrate the resulting ontology, we provide Figure 1, which represents all its main concepts (norms, agents, values, outcomes and statements) and their high-level relationships. The more specific Figure 2 represents the key notions of the ontology and their relationships with higher detail. For more detail, please refer to the following Github repository[9].

The VAE ontology (core module)[10] consists of 1575 axioms, 129 classes, 121 object properties and 5 datatype properties. Also, it features 7 SWRL rules for binary relation properties such as transitivity or reflexivity. Most of the axiom complexity is due to the DOLCE+DnS Ultralite (Borgo et al., 2022) imported classes. DOLCE has a very detailed object property and class hierarchy (1549 axioms in total) that allows to represent both specific and abstract notions.

[6] https://oeg.fi.upm.es/files/pdf/NeOnGlossary.pdf

[7] We treat values as mere "abstract concepts", in line with ValueNet's (De Giorgis et al., 2022) notion.

[8] The assumed argumentation theory comes from argument mining proposals (Lawrence and Reed, 2019; Segura-Tinoco et al., 2022) where *claims* and *premises* are the basic argumentative units linked via *argument relations*. Similarly, we consider arguments as *statements* composed by premises and claims, that are related via certain *criteria*.

[9] https://github.com/andresh26-uam/vae-ontology

[10] VAE ontology core: https://w3id.org/def/vaeontology

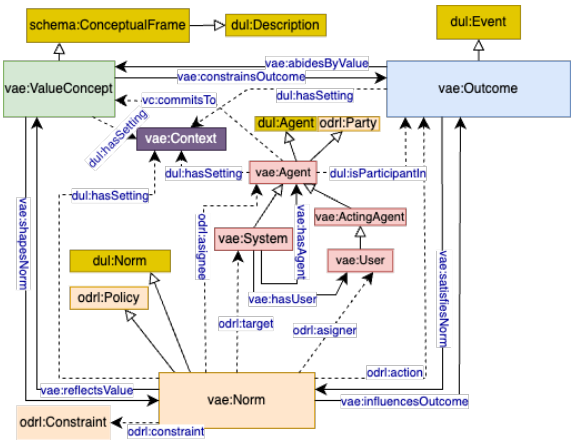

Figure 2: The centre of the VAE ontology, featuring *agents*, *values*, *norms* and *outcomes*. Most of the terminology is rooted in (Steels, 2023), among others.

4 CASE STUDIES

In this section we are interested in evaluating the representation power of the ontology to formalize VAE proposals in agent-based value-aware (normative) systems. To do so we explicitly conceptualize and implement three inluential VAE proposals (Montes and Sierra, 2022; Osman and d'Inverno, 2023; Serramia et al., 2018) in the next sections.

The methodology followed to implement into ontologies those theories involved again NeOn methodology activities, following a similar process to Section 3. In short, we reused the VAE core, coded the necessary new notions for each case and then populated the ontologies with individuals from examples or use cases used in the corresponding papers.

For each new ontology, we present its representation power by giving a mapping from each notion present in the corresponding theory to classes and properties from the ontology (adding its relation to the VAE ontology core). Then, we discuss the conciseness of the representation, accounting to the *number of axioms per notion* and *classes per notion* ratios[11]. The notions of conciseness are inspired by (Davis et al., 1993).

4.1 Case 1: Synthesis and Properties of Optimally Value-Aligned Normative Systems (Montes and Sierra, 2022)

The problem presented at Montes and Sierra (Montes and Sierra, 2022) consisted of finding parametrizations of *parametric norms* (composing *normative systems*) that, applied to a certain *MAS system* (Multi-Agent System), are "*optimally aligned*" with a set of values. The authors assumed a consequentialist view, i.e. that the alignment of a concrete parametrization of a set of norms relies only on the alignment with values of the possible evolutions of the MAS, which should be evaluated via the *semantics of the values* in the *states* that are allowed to happen by the norms.

Most of the terminology is present in the core module, though some concepts require a specific representation (e.g. *parametric norms*). Table 1 presents how these notions have been introduced in the ontology framework.

In total, taking into account the referenced classes from the VAE ontology, the new ontology[12] and the modelled example actively use 800 axioms, 90 individuals, 50 classes, 53 object properties and 6 datatype properties. The average number of axioms utilized per required notion is $\frac{800}{8} = 100$. It must be noted that most of the axiomatization effort was put in technical details of the paper's running example, rather than on actually new VAE concepts. Thus, the classes per notion ratio is high at 6.25, but if we remove the example-dependent classes we get 39 classes, so the ratio drops to 4.875.

4.2 Case 2: A Computational Framework of Human Values for Ethical AI (Osman and d'Inverno, 2023)

This work presents an original "computational definition of values" based on taxonomical representations of values without specifically committing to a consequentialist or deontological view.

The theory assumes that values are general concepts that become more specific as one travels down the taxonomy, becoming "concrete and computational" at leaf nodes (i.e. becoming *properties* that have a computable *degree of satisfaction* at the world *states*"). Also, the authors define the importance of the concepts and properties building the taxonomy, which is fully *context-dependent*. Finally, they consider a characterization of *value-alignment*, proposing a user-defined function that calculates the alignment of an entity's behaviour depending on the context.

Table 2 presents the specific notions integrated in the ontology[13] as classes, while Table 3 does the same for notions integrated as OWL properties (`owl:DatatypeProperty`/`owl:ObjectProperty`).

[11]Calculated with Protégé (v5.6.3) after importing the ontology and removing the core module import statement.

[12]https://w3id.org/def/vaeontology_montes_sierra

[13]https://w3id.org/def/vaeontology_osman_dInverno

Table 1: Represented notions from (Montes and Sierra, 2022) in the VAE ontology as OWL classes and some important axioms written in OWL Manchester Syntax. The prefix `ms:` is used to identify terms of the new ontology. The $\sqsubseteq$ is used to denote inheritance.

Notion in (Montes and Sierra, 2022)	**Ontology class + (relevant classification)**
Norm, Action, State, Transition, Agent, MAS	`vae:Norm`, `vae:Action`, `vae:State`, `vae:Transition`, `vae:Agent`, `vae:System`.
Parametric norm	`ms:ParametricNorm` ($\sqsubseteq$ `vae:Norm`)
Normative System	`ms:NormativeSystem` ($\sqsubseteq$ `vae:Norm`)
Path (with final state)	`vae:Path` $\cap$ `:hasOutState some vae:State`
Norm Parameter	`dul:Parameter`
Semantics Function	`ms:SemanticsFunction`
Aggregation Function (of Semantics Functions)	`ms:SemanticsFunctionAggregation` ($\sqsubseteq$ `vae:QuantitativeVaeProperty`)
Normative System Alignment	`ms:NormativeSystemAlignment` ($\sqsubseteq$ `vae:QuantitativeVaeProperty`). Added axioms to indicate that it is measured on a set of possible paths after applying a normative system (i.e. over $\mathcal{P}^N$ from (Montes and Sierra, 2022)).
Optimal Normative System Alignment	`ms:OptimalNormativeSystemAlignment` ($\sqsubseteq$ `ms:NormativeSystemAlignment` $\cap$ `vae:OptimizedProperty`).

Note that some notions were already modelled in the core module, e.g. the notions of *context* and *properties*. Of course, more OWL axioms and SWRL rules were added, for instance, to maintain the direct acyclic graph (DAG) structure of the taxonomies, respect of importance *condomains*, and propagate information.

We considered as use cases the example taxonomies present in Figures 1, 2 and 3 from the paper, that represent different taxonomies, and automatically calculated their alignment function values.

In total, taking into account the referenced classes from the VAE ontology, this case ontology actively uses 734 axioms, 90 individuals, 46 classes, 45 object properties and 10 datatype properties. That accounts for an axioms per notion ratio of $\frac{734}{14} = 52.43$, suggesting a better core ontology reuse than in the last case. The classes per notion ratio is also lower at 3.28.

4.3 Case 3: Moral Values in Norm Decision Making (Serramia et al., 2018)

The third proposal (Serramia et al., 2018) approached a similar problem to Montes and Sierra's, namely, finding the subset of norms—*norm system*—with maximum value support (considering also its representation power and minimum implementation cost) from a set of feasible norms—*norm net*—. The solution is obtained by solving a linear optimization problem. The main difference from Montes and Sierra's is that Serramia and colleagues define the relation between norms and values assuming a deontological stance, defining a *support rate function* that characterizes the degrees of promotion or demotion of some values by one norm. The theory takes into account rich relationships between norms (*exclusivity*, *substitutability*, *generalization*) and values (*value systems*).

The notions that were representable in the ontology are given at Table 4 (classes) and Table 5 (properties). This case extensively uses SWRL rules, so we recommend the reader to inspect the full ontology[14].

Serramia's theoretical approach requires utilizing most of the VAE ontology, such as pairwise relations and comparisons between both norms and values (`vae:ComparisonStatement`) and quantitative properties measurable in norms, that are must be defined as instances of deontic operators, e.g. `odrl:Permission`.

New notions were to be defined, though, namely, statements about the new norm binary relations (modelled as context-based classes) and optimization problems. And although value systems are again presented as a DAG (similarly to the case in Section 4.2), for parallelisms with the norm representation, the DAG structure of each value system was implemented with SWRL rules over pairwise comparisons.

As proof of concept, we modelled the main examples from the paper, e.g. Examples 2.1 (a basic norm net with different *agents*, *norms* and their binary relations), 4.1 (a sample *value system*) and 4.2 (presenting the optimization results, and the inferred preferences of norms and norm systems based on the value preferences).

This case ontology uses 671 axioms, 55 individuals, 80 classes, 64 object properties and 4 datatype properties, for a total of 27 notions (22 translated into classes, 5 into properties). The number of new SWRL

[14]The ontology about (Serramia et al., 2018): https://w3id.org/def/vaeontology_moral_values_in_norm_DM

Table 2: Represented notions from (Osman and d'Inverno, 2023) in the VAE ontology as OWL classes and some important axioms. The prefix `odi:` is used to identify terms of the new ontology.

Notion in (Osman and d'Inverno, 2023)	**Ontology class + (relevant classification)**
State, Agent, System, Context	`vae:State`, `vae:Agent`, `vae:System`, `vae:Context`. Basic terminology.
Context-based Value Taxonomy	`odi:ValueTaxonomyStatement` ($\sqsubseteq$ `vae:AgentStatement` $\cap$ $\exists$ `dul:hasSetting`. `vae:Context`)
Nodes in a value taxonomy	`odi:TaxonomyNode` ($\sqsubseteq$ `vae:AgentStatement`)
Label nodes ("representing abstract value *concepts*")	`odi:ConceptNode` ($\sqsubseteq$ `odi:TaxonomyNode`)
Property nodes	`odi:PropertyNode` ($\sqsubseteq$ `odi:TaxonomyNode`)
Properties verified in states	`odi:TaxonomyProperty` ($\sqsubseteq$ `odi:QuantitativeVaeProperty`)
Importance of a Node	`odi:NodeImportance` ($\sqsubseteq$ `vae:QuantitativeVaeProperty`)
Aggregation of importance function	`odi:AggregationOfImportance` ($\sqsubseteq$ `vae:QuantitativeVaeProperty`). Stands for the calculation of importance of a Taxonomy.
Condomain	`dul:Region`
Alignment function	`odi:TaxonomyAlignment` ($\sqsubseteq$ `vae:ValueProperty` $\cap$ `vae:QuantitativeVaeProperty` $\cap$ `vae:AggregationFunction`)

Table 3: Notions from (Osman and d'Inverno, 2023) in the VAE ontology as OWL object and datatype properties. The prefix `odi:` is used to identify terms of the new ontology.

Notion in (Osman and d'Inverno, 2023)	**Ontology property + (relevant classification)**
Concept/Property generalization	`odi:directlyGeneralizesNode` ($\sqsubseteq$ `odi:generalizesNode`)
Condomain of Taxonomy	`odi:hasCondomain` ($\sqsubseteq$ `dul:hasRegion`)
Degree of satisfaction	`odi:degreeOfSatisfaction` ($\sqsubseteq$ `dul:hasDataValue`)
Importance of a node	`odi:importanceValue` ($\sqsubseteq$ `dul:hasDataValue`)

rules is 16. That accounts for an axioms per notion average of $\frac{671}{27} = 24.85$, halving the ratio of the last case. This is due to an extensive reuse of the VAE ontology axioms, and having more (overlapping) notions. If we look at the classes per notion ratio we get a similar one as the previous case, 2.96.

5 DISCUSSION

The three case proposals where successfully implemented with competent coverage. In the Case 4.1 (8 new notions) we represented the fundamental theory for representing optimally-aligned normative systems and the evolution based on sequences of transitions leaving out of scope the second part of the paper about model analysis. In Case 4.2 (14 new notions), we implemented all the logic for consistently building context-dependent value taxonomies with the notions of importance and alignment (using 14 notions). In Case 4.3 (27 new notions) we managed to control the compatibility of the inserted individuals within the theory by checking sound norm systems properties; representation and inference of norm relations; and selecting the value preferences of a value system that respect the desired DAG structures.

Limitations of the ontological representations are most due to OWL+SWRL limited representation and inference power, sometimes limited by the Open World Assumption. For example, it was not always possible to calculate the aggregation of numerical values (e.g. Monte-Carlo estimation of the `ms:NormativeSystemAlignment` in Case 4.1, alignment function and aggregation of importance in Case 4.2 or value preference utilities in Case 4.3) nor provide inferences via second order logic and negation (e.g. impossibility to infer what norm systems are conflict-free or non-redundant in Case 4.3 or to define *monotonocity* and *idempotence* in Case 4.2).

In general, we highlight the fact that the ontologies remain interoperable even them assuming opposed views such as consequentialism (Case 4.1) or deontology (Case 4.2). Also, we highlight the increasing metrics of conciseness achieved despite the increasing notion complexity of the cases seen.

6 CONCLUSIONS

In this paper we presented a new ontology for value-aware agent-based systems. It aims to be a step towards a common representation for key concepts in the emerging field of *value awareness engineering* (VAE), comprising a compilation of computational interpretations of social science definitions, thereby supporting the research community by easing the implementation gap for new value-aware systems. The ontology was implemented in OWL, using SWRL to enhance its representation power, and following

Table 4: Represented notions from (Serramia et al., 2018) in the VAE ontology as OWL classes and some important axioms. The prefix `mvndm:` is used to identify terms of the new ontology.

Notion in (Serramia et al., 2018)	**Ontology class + (relevant classification)**
Norm, Agent	`mvndm:Norm` ($\sqsubseteq$ `vae:Norm` $\cap$ `odrl:Rule`), `vae:Agent`.
Norm Exclusivity, Substitutability, Direct Generalisation	`mvndm:Exclusivity`, `mvndm:Substitutability`, `mvndm:DirectGeneralizationStatement` ($\sqsubseteq$ `vae:RelationStatement`)
Indirect generalization of a norm	`mvndm:TransitiveGeneralizationStatement` ($\sqsubseteq$ `vae:RelationStatement`).
Norm system	`mvndm:NormSystem` ($\sqsubseteq$ `vae:Norm` $\cap$ `dul:Collection`)
Conflict-free norm system, Non-redundant norm system	`mvndm:ConflictFreeNormSystem`,`mvndm:NonRedundantNormSystem` ($\sqsubseteq$ `mvndm:NormSystem`)
Sound norm system	`mvndm:SoundNormSystem` ($\equiv$ `mvndm:ConflictFreeNormSystem` $\cap$ `mvndm:NonRedundantNormSystem`)
Norm Net	`mvndm:NormNet` ($\sqsubseteq$ `vae:AgentStatement` $\cap$ `dul:Collection`)
Norm cost, representation Power	`mvndm:NormCost`, `mvndm:NormRepresentationPower` ($\sqsubseteq$ `vae:QuantitativeVaeProperty` $\cap$ (≥ 1)`.vae:measuredOnConditionedEntity mvndm:Norm`)
Norm system cost, Rep. Power	(analogous to last notion)
Maximum Norm System Problem	`mvndm:MaximumNormSystemProblem` ($\sqsubseteq$ `vae:VaeStatement`)
Value System	`mvndm:ValueSystem` ($\sqsubseteq$ `vae:AgentStatement` $\cap$ `dul:Collection`)
Partial order of value preferences	`mvndm:PartialOrderValueComparison` ($\sqsubseteq$ `vae:ValueComparisonStatement` $\cap$ `vae:TransitiveRelationStatement`)
Support rate function	`mvndm:SupportRateComponent` ($\sqsubseteq$ `vae:QuantitativePromotionDemotion`)
Value preference utility	`mvndm:ValuePreferenceUtility` ($\sqsubseteq$ `vae:QuantitativeVaeProperty`)
Value support	`mvndm:NormValueSupport` ($\sqsubseteq$ `vae:QuantitativeVaeProperty`)
Norm system preference relation	`mvndm:NormComparisonStatement` ($\sqsubseteq$ `vae:ComparisonStatement`)
Value-based norm optimisation problem	`mvndm:ValueBasedNormOptimizationProblem` ($\sqsubseteq$ `mvndm:MaximumNormSystemProblem`)

Table 5: Notions from (Serramia et al., 2018) in the VAE ontology as OWL properties. The prefix `mvndm:` is used to identify terms of the new ontology.

Notion in (Serramia et al., 2018)	**Ontology property + (relevant classification)**
Budget	`mvndm:hasBudget`
Norm/Value Comparison	`vae:comparisonHasSuperior` `vae:comparisonHasInferior`
Utility in a comparison of norms/values	`vae:hasPropertyOfSuperior`/ `vae:hasPropertyOfInferior` ($\sqsubseteq$ `vae:expressesProperty`)
Norm system of a norm net	`mvndm:isSubsetOfNormNet` ($\sqsubseteq$ `dul:isMemberOf`)
DAG preservation	`mvndm:isDiscardedForVS`/ `mvndm:isNotDiscardedForVS`

the NeOn methodology; thus, conveying to stablished standards for ontology engineering.

The expressive power of the ontology in relation to value-aware systems was illustrated through case studies comprising the representation of three influential theories from the VAE field as well as their main running examples. We achieved concise yet deep representations of the proposals, integrated without logical inconsistencies despite their diverse philosophical grounding.

This work opens up several lines of future work. Firstly, we will look into an implementation for the argumentative framework (which remains at the representation level). Secondly, the use of SHACL[15] for constraint validation with closed-world assumptions to enhance the expressive power of the ontology needs to be explored. Finally, the interoperability facet of the ontology is to be tested in a simulated or deployed value-aware system.

ACKNOWLEDGEMENTS

This work has been supported by grant VAE: TED2021-131295B-C33 funded by MCIN/AEI/ 10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR", by grant COSASS: PID2021-123673OB-C32 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", and by the AGROBOTS Project of Universidad Rey Juan Carlos funded by the Community of Madrid, Spain.

REFERENCES

Arnold, T., Kasenberg, D., and Scheutz, M. (2017). Value alignment or misalignment – what will keep systems accountable? In *AAAI Workshop on AI, Ethics, and Society*.

[15]https://www.w3.org/TR/shacl/

Balakrishnan, A., Bouneffouf, D., Mattei, N., and Rossi, F. (2019). Incorporating behavioral constraints in online ai systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3–11.

Bench-Capon, T., Atkinson, K., and McBurney, P. (2012). Using argumentation to model agent decision making in economic experiments. *Autonomous Agents and Multi-Agent Systems*, 25:183–208.

Borgo, S., Ferrario, R., Gangemi, A., Guarino, N., Masolo, C., Porello, D., Sanfilippo, E. M., and Vieu, L. (2022). DOLCE: A descriptive ontology for linguistic and cognitive engineering. *Applied Ontology*, 17(1):45–69.

Chisholm, R. M. (1963). Supererogation and offence: A conceptual scheme for ethics. *Ratio (Misc.)*, 5(1):1.

Davis, A., Overmyer, S., Jordan, K., Caruso, J., Dandashi, F., Dinh, A., Kincaid, G., Ledeboer, G., Reynolds, P., Sitaram, P., Ta, A., and Theofanos, M. (1993). Identifying and measuring quality in a software requirements specification. In *Proceedings First International Software Metrics Symposium*, pages 141–152.

De Giorgis, S., Gangemi, A., and Damiano, R. (2022). Basic human values and moral foundations theory in valuenet ontology. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 3–18. Springer.

Fornara, N. and Colombetti, M. (2010). Ontology and time evolution of obligations and prohibitions using semantic web technology. *Lecture Notes in Computer Science*, 5948 LNAI:101 – 118.

Gangemi, A. (2008). Norms and plans as unification criteria for social collectives. *Autonomous Agents and Multi-Agent Systems*, 17(1):70–112.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2003). Sweetening wordnet with dolce. *AI magazine*, 24(3):13–13.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of *Advances in Experimental Social Psychology*, pages 55–130. Academic Press.

Grosof, B. N., Horrocks, I., Volz, R., and Decker, S. (2003). Description logic programs: Combining logic programs with description logic. In *Proceedings of the 12th international conference on World Wide Web*, pages 48–57.

Holgado-Sánchez, A., Arias, J., Moreno-Rebato, M., and Ossowski, S. (2023). On admissible behaviours for goal-oriented decision-making of value-aware agents. In *Multi-Agent Systems*, pages 415–424, Cham. Springer Nature Switzerland.

Ianella, R. and Villata, S. (2018). ODRL information model 2.2. W3C Recommendation, W3C.

Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Lera-Leri, R., Bistaffa, F., Serramia, M., Lopez-Sanchez, M., and Rodriguez-Aguilar, J. (2022). Towards pluralistic value alignment: Aggregating value systems through lp-regression. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 780–788. IFAAMAS.

Montes, N., Osman, N., Sierra, C., and Slavkovik, M. (2023). Value engineering for autonomous agents. *CoRR*, abs/2302.08759.

Montes, N. and Sierra, C. (2021). Value-guided synthesis of parametric normative systems. pages 907–915. IFAAMAS.

Montes, N. and Sierra, C. (2022). Synthesis and properties of optimally value-aligned normative systems. *Journal of Artificial Intelligence Research*, 74:1739–1774.

Osman, N. and d'Inverno, M. (2023). A computational framework of human values for ethical ai.

Poole, D. L. and Mackworth, A. K. (2010). *Artificial Intelligence: foundations of computational agents*. Cambridge University Press.

Poveda-Villalón, M., Gómez-Pérez, A., and Suárez-Figueroa, M. C. (2014). Oops! (ontology pitfall scanner!): An on-line tool for ontology evaluation. *Int. J. Semantic Web Inf. Syst.*, 10:7–34.

Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., and Rodriguez-Aguilar, J. A. (2022). Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24:9.

Russell, S. (2022). *Artificial Intelligence and the Problem of Control*, pages 19–24. Springer International Publishing, Cham.

Schwartz, S. H. (1992). 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.

Segura-Tinoco, A., Holgado-Sánchez, A., Cantador, I., Cortés-Cediel, M., and Bolívar, M. R. (2022). A conversational agent for argument-driven e-participation.

Serramia, M., Lopez-Sanchez, M., and Rodriguez-Aguilar, J. A. (2020). A qualitative approach to composing value-aligned norm systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 1233–1241, Richland, SC. IFAAMAS.

Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Rodriguez, M., Wooldridge, M., Morales, J., and Ansotegui, C. (2018). Moral values in norm decision making. *IFAAMAS*, 9.

Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., and Perelló, A. (2021). Value alignment: a formal approach. *CoRR*, abs/2110.09240.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, 5(2):51–53. Software Engineering and the Semantic Web.

Soares, N. (2018). The value learning problem. *Artificial Intelligence Safety and Security*.

Steels, L. (2023). Values, norms and ai – introduction to the vale workshop. In *Pre-proceedings of the ECAI Workshop on Value Engineering (VALE)*, page 6–8.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2015). The neon methodology framework: A scenario-based methodology for ontology development. *Applied Ontology*, 10(2):107–145.