An Architecture for Learning Context-based Value Systems

Andrés Holgado-Sánchez, Sascha Ossowski, Holger Billhardt

CETINIA, Universidad Rey Juan Carlos
Tulipán St. (Unnumbered)
Móstoles, Madrid 28939 Spain
Corresponding author email: andres.holgado@urjc.es

Abstract

As cyberphysical systems become increasingly autonomous, it must be assured that their behaviour is aligned with human values. Insights from Social Science reveal that human value systems depend on context. In this *extended abstract* we introduce a representation for contextual value systems and propose a model for learning them from examples.

Introduction

The value alignment problem (Russell 2022) states that decisions taken by AI systems must obey ethical principles and human values. A way to achieve this is to endow them with explicit representations of human values and value reasoning abilities. *Context dependency* is key in such *value-aware* AI systems (Osman and d'Inverno 2024): e.g., the importance of the values of "privacy" and "efficiency" may vary when evaluating the same alternative in different contexts.

Given the difficulty of modelling values manually, approaches like Axies (Liscio et al. 2021) and moral value detection (Rink, Lobachev, and Vorontsov 2024) apply context-sensitive value learning through text analysis but cannot reason about value preferences. The work by (Holgado-Sánchez et al. 2025) can infer value representations and value systems (value preferences) as reward functions using inverse reinforcement learning (Ng and Russell 2000), but does not account for context dependency in these representations. The contribution of this paper is twofold: it extends the aforementioned framework with context-sensitivity, and presents the architecture of a computational framework for context-based value learning.

Representing context-based value systems

We set out from a set of m values $V = \{v_1, \dots, v_m\}$ in an environment with a set of decision alternatives called *entities* E. Each entity $e \in E$ is described by a feature vector $\boldsymbol{\phi}(e) \in \Phi$. For example, values that guide decision-making in route choice (Prato 2009) may include *ecology* and *security*, entities to choose from are the alternative routes, and features include *route length* or *average speed*. A value $v \in V$ acquires a particular meaning by *grounding* it in environment features. *Value alignment functions* allow for computing the level of alignment of alternatives with values:

Definition 1 (Value alignment function/Grounding) *Given a value v, the function $\mathcal{A}_v : E \to \mathbb{R}$ is a **value alignment function** for v. Then, the **grounding** for V, $G_V : E \to \mathbb{R}^m$, is $G_V = (\mathcal{A}_{v_1}, \dots, \mathcal{A}_{v_m})$ in E.*

In our route choice example, the grounding of the value *ecology* may refer to the expected CO_2 consumption of a certain route e, that can be estimated from the features *route length* and *average speed*. While we assume that groundings are socially-agreed upon in a society, each agent j may hold its own *value system*. We model the latter as a preference relation $\preccurlyeq_{j,G_V}$ over entities in E based on a grounding G_V. Furthermore, we characterise a context c by a series of features $\psi(c) \in \Psi$ that influence agent preferences. E.g, a context may comprise features that determine whether a route is to be chosen for a leisure or for a business trip.

Definition 2 (Contextual value system) *Let G_V a grounding. The **contextual value system** of an agent j is a mapping from contexts to value systems: $VS_{j,G_V}(c) = \left\{\preccurlyeq^c_{j,G_V}\right\}$.*

Contextual value system functions are defined by importance weights for each value v_i based on which, for each context, the degrees of alignment of an entity e with v_i are linearly aggregated, so as to represent the corresponding value system.

Definition 3 (Contextual value system function) *Let VS_{j,G_V} be the contextual value system of agent j. The function $\mathcal{A}_{j,G_V} : C \times E \to \mathbb{R}$ is the **contextual value system function** of j assuming the grounding G_V, defined as $\mathcal{A}_{j,G_V}(c, e) = W^c_j \cdot G_V(e)$, where $W^c_j = (w^{v_1}_j(c), \dots, w^{v_m}_j(c))$ represents the value weights reflecting j's value system at context $c \in C$.*

As in previous work, we assume that the weights W^c_j are restricted to $[0, 1]^m$ and that their sum across values is 1, so that agents cannot be opposed to the promotion of a value or be completely unaware of all of them.

With this representation model we can redefine the *value system learning* problem (Holgado-Sánchez et al. 2025), adapting for context-dependency.

Definition 4 (Contextual value system learning) *v The **contextual value system learning** problem consists of (i)*

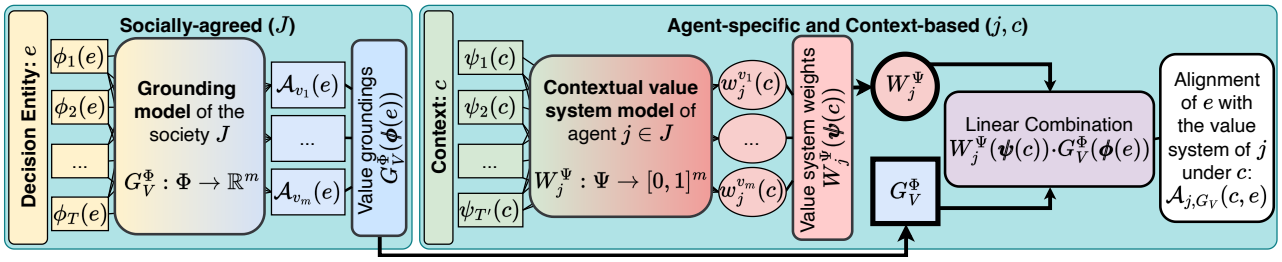

Figure 1: The proposed architecture applicable to learning agent-specific and context-based value systems.

inferring a socially-agreed grounding $G_V : E \to \mathbb{R}^m$ *that estimates value alignment, and (ii) learning a contextual value system function* $\mathcal{A}_{j,G_V} : C \times E \to \mathbb{R}$ *for each agent* $j \in J$ *based on* G_V*, which reduces to finding the corresponding contextual value system weights* $W_j^c \in [0,1]^m$.

Contextual value system learning architecture

In this section, we describe our second contribution: an architecture for contextual value system learning. It consists of two modules. The first module (Figure 1, left) features the model $G_V^\Phi : \Phi \to \mathbb{R}^m$ that represents the socially-agreed grounding G_V and is based on the entity features Φ. The second module (right) is agent-specific (for each $j \in J$) and also context-dependent, as it includes the learnable model $W_j^\Psi : \Psi \to [0,1]^m$ that estimates the weights W_j^c from the features of the context c. Then, with the linear combination of the weights and the groundings, an estimation of the contextual value system alignment function of the agent is derived: $\mathcal{A}_{j,G_V}(c,e) = W_j^\Psi(\boldsymbol{\psi}(c)) \cdot G_V^\Phi(\boldsymbol{\phi}(e))$.

While the context feature space can be infinite or dense, the space of possible contextual value systems should not, at least for a human. E.g., if the time of day is a context feature, W_j^Ψ might assign different weights at 9:00 AM than at 9:01 AM. To prevent these unreasonable predictions, we suggest limiting the number of contexts that trigger different value systems and, if possible, make the relationship between contexts and value system interpretable. To do so, we suggest implementing clustering (Chakraborty et al. 2024), rule-based (Veronese et al. 2024) or meta-heuristic (Bruns, Dunkel, and Seremet 2023) approaches in the contextual value system model architecture.

Given the architecture, we propose a learning setting to ground Definition 4. Let an environment in which for each agent $j \in J$ we observe a dataset $\mathcal{D}_j = \{(e_k, e'_k, y_k^j, c_k)\}_{k=1}^K$, where $y_i^j \in [0,1]$ captures j's relative preference for entity e_i over e'_i at context c_i according to j's value system. Consider also another dataset from domain experts, namely, $\mathcal{D}_V = \{(e_l, e'_l, y_l^{v_1}, \ldots, y_l^{v_2})\}_{l=1}^L$ where $y^{v_k} \in [0,1]$[1] is the annotated degree of preference of e_i over e'_i regarding their alignment with value v_k.

[1]Though they are defined generally as quantitative measures in $[0,1]$, both $y_l^{v_k}$ and y_i^j may represent qualitative preferences if they are restricted to $\{0, 1.0, 0.5\}$, to indicate that e_i is preferred over e'_i, the inverse relation, or indifference, respectively.

Additionally, let $\mathcal{L}_{G_V^\Phi}(\mathcal{D}_V)$ and $\mathcal{L}_{W_j^\Psi | G_V^\Phi}(\mathcal{D}_j)$ be two *loss* functions. The first measures the error of approximating y^{v_k} from $G_V^\Phi(e)$, and the second, the error of approximating y^j from $\mathcal{A}_{j,G_V}(c,e) = W_j^\Psi(\boldsymbol{\psi}(c)) \cdot G_V^\Phi(\boldsymbol{\phi}(e))$. Minimizing these two losses in a structured way (Definition 5), we obtain optimal models $(G_V^\Phi)^*$ and $(W_j^\Psi)^*$ that maximally represent the preferences in the datasets, solving the two tasks from Definition 4. To define these losses, inspiration can be taken from RLHF (Christiano et al. 2017).

Definition 5 *The* ***contextual value system learning problem*** *consists of solving the bi-level optimization problem:*

$$(W_j^\Psi)^* = \arg\min_{W_j^\Psi} \sum_{j \in J} \mathcal{L}_{W_j^\Psi | (G_V^\Phi)^*}(\mathcal{D}_j)$$

$$\textit{subject to} \quad (G_V^\Phi)^* \in \arg\min_{\theta_V} \mathcal{L}_{G_V^\Phi}(\mathcal{D}_V).$$

Approaching this bi-level optimization problem requires prioritizing approximating the value dataset $\mathcal{D}_V$ over the value system ones $\mathcal{D}_j$. To do that, and assuming differentiable loss functions and learning our models through gradient descent, we can use a weighting factor $\lambda > 0$ to treat the loss $\mathcal{L}_{G_V^\phi}$ as a soft constraint and minimize a global loss $\mathcal{L}(\mathcal{D}_j, \mathcal{D}_V) = \sum_{j \in J} \mathcal{L}_{W_j^\Psi | (G_V^\Phi)^*}(\mathcal{D}_j) + \lambda \mathcal{L}_{G_V^\Phi}(\mathcal{D}_V)$. It is also possible to estimate a suitable value for λ during optimization (Cotter, Jiang, and Sridharan 2019). An alternative to differentiable approaches could be techniques from the field of operations research, employed in the related problems of finding maximally-aligned normative systems (Serramia et al. 2018) and aggregating value-based preferences of multiple agents (Lera-Leri et al. 2022).

Applicability and future work

The proposed approach will allow us to estimate the value system of new agents, recommend alternatives for custom preference weights, or predict agent preferences in simulated contexts. Potential application domains include route choice (Zhao and Liang 2023) or value-aware recommender systems (De Biasio et al. 2023).

In future work, we will implement and evaluate the proposed architecture with real-world datasets on route choice, extending synthetic experiments in (Holgado-Sánchez et al. 2025). This will require extending the context model to sequential decision-making. Other challenges include the recognition of trends in the diverse value systems of a full society of agents, and support for heterogeneous value groundings (context or agent-based).

Acknowledgments

This work is 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

Bruns, R.; Dunkel, J.; and Seremet, S. 2023. Learning Ship Activity Patterns in Maritime Data Streams: Enhancing CEP Rule Learning by Temporal and Spatial Relations and Domain-Specific Functions. *IEEE Transactions on Intelligent Transportation Systems*, 24(10): 11384–11395.

Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Manocha, D.; Huang, F.; Bedi, A.; and Wang, M. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *Proc. 41st Int. Conf. on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 6116–6135. PMLR.

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Proc. NIPS'17*, 4302–4310.

Cotter, A.; Jiang, H.; and Sridharan, K. 2019. Two-Player Games for Efficient Non-Convex Constrained Optimization. In Garivier, A.; and Kale, S., eds., *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, 300–332. PMLR.

De Biasio, A.; Montagna, A.; Aiolli, F.; and Navarin, N. 2023. A systematic review of value-aware recommender systems. *Expert Systems with Applications*, 226: 120131.

Holgado-Sánchez, A.; Bajo, J.; Billhardt, H.; Ossowski, S.; and Arias, J. 2025. Value Learning for Value-Aligned Route Choice Modeling via Inverse Reinforcement Learning. In Osman, N.; and Steels, L., eds., *Value Engineering in Artificial Intelligence*, 40–60. Cham: Springer Nature Switzerland. ISBN 978-3-031-85463-7.

Lera-Leri, R.; Bistaffa, F.; Serramia, M.; Lopez-Sanchez, M.; and Rodriguez-Aguilar, J. A. 2022. Towards pluralistic value alignment: Aggregating value systems through lp-regression. *openaccess.city.ac.ukR Lera-Leri, F Bistaffa, M Serramia, M Lopez-Sanchez, J Rodriguez-AguilarProceedings of the 21st International Conference on Autonomous, 2022•openaccess.city.ac.uk*, 9: 780–788.

Liscio, E.; van der Meer, M.; Cavalcante Siebert, L.; Mouter, N.; Jonker, C.; and Murukannaiah, P. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proc. AAMAS'21*, 799–808.

Ng, A. Y.; and Russell, S. J. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, 663–670. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Osman, N.; and d'Inverno, M. 2024. A Computational Framework of Human Values. In *Proc. AAMAS'24*, 1531–1539.

Prato, C. G. 2009. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2(1): 65–100.

Rink, O.; Lobachev, V.; and Vorontsov, K. 2024. Detecting Human Values and Sentiments in Large Text Collections with a Context-Dependent Information Markup: A Methodology and Math. In *Proc. HCII'24*, 372 – 383.

Russell, S. 2022. Artificial Intelligence and the Problem of Control. In Werthner, H.; Prem, E.; Lee, E. A.; and Ghezzi, C., eds., *Perspectives on Digital Humanism*, 19–24. Springer.

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.

Veronese, C.; Meli, D.; Bistaffa, F.; Rodríguez-Soto, M.; Farinelli, A.; and Rodríguez-Aguilar, J. A. 2024. Inductive Logic Programming for Transparent Alignment with Multiple Moral Values. In *CEUR workshop proceedings*, volume 7, 84–88. ISBN 9772081415.

Zhao, Z.; and Liang, Y. 2023. A deep inverse reinforcement learning approach to route choice modeling with context-dependent rewards. *Transportation Research Part C: Emerging Technologies*, 149: 104079.