Value awareness and process automation: a reflection through school place allocation models

Joaquín Arias[1], Mar Moreno-Rebato[1], Jose A. Rodriguez-García[1] and Sascha Ossowski[1]

[1]*CETINIA, Universidad Rey Juan Carlos, Madrid, Spain*

Abstract

The proposed regulatory framework for artificial intelligence and the EU General Data Protection Regulation make it necessary for automated reasoners to *justify* their conclusions in human-understandable terms. In addition, there are ethical and legal concerns that should be addressed to ensure that the advice given by such AI systems is aligned with human values.

Value-aware systems address this challenge by explicitly representing and reasoning with norms and values applicable to a problem domain. In procedures of the public administration, for instance, such systems may provide support to decision-makers and, ultimately, enable the automation of (part of) these administrative processes. However, this requires the capability to determine as to how far a particular legal model is aligned with a certain value system. The s(LAW) legal reasoner based on Answer Set Programming has proven capable of adequately modelling administrative processes with discretion.

This article is an extended abstract of a work in progress where we analyse two (political) strategies for school place allocation in educational institutions supported with public funds, that differently weigh values such as equality, fairness, and non-segregation. We plan to illustrate how s(LAW) models these scenarios, and how automated reasoning with these models can answer various questions regarding their value-alignment.

Keywords

Legal reasoning, Goal-directed execution, Answer set programming, Inductive logic programming

The automation of all sorts of processes through Artificial Intelligence systems has made significant progress over the last years. More recently, it has become apparent that this development needs to be accompanied by means that guarantee, as much as possible, the protection of the people that are affected by the decisions generated by such systems. Whether though self-regulation or soft law (guides, guidelines, codes of conduct, declarations, ethical charters on AI) or through legal regulation (GDPR and proposed EU Regulation on AI), interest and concern has increased for safeguarding the fundamental rights and safety of people affected by AI systems. Promoting a reliable AI, focused on the human being, is of foremost importance because, even though the designers' intentions are good, autonomous AI systems may cause significant harm.

To this respect, the novel field of value-awareness engineering [1] is emerging, which claims

3rd Workshop on Goal-directed Execution of Answer Set Programs (GDE'23), July 10, 2023

joaquin.arias@urjc.es (J. Arias); mar.rebato@urjc.es (M. Moreno-Rebato); joseantonio.rodriguez@urjc.es (J. A. Rodriguez-García); sascha.ossowski@urjc.es (S. Ossowski)

0000-0003-4148-311X (J. Arias); 0000-0002-4177-9239 (M. Moreno-Rebato); 0000-0002-6362-9880 (J. A. Rodriguez-García); 0000-0003-2483-9508 (S. Ossowski)

that it is possible to formally represent values, and to reason with and about them, paving the way for future *machine morality*. In fact, current AI systems are not value-aware. In one recent example, GPT3, developed by OpenAI, encouraged a person to commit suicide[1] and offered help on how to do so, violating the fundamental human value of not encouraging harm. In this context, it is necessary to elicit, model and aggregate the human values that a given community may (or may not) collectively agree upon, so that we can apply simulation, reasoning or learning-based techniques, e.g. to account for value-driven decision making [2], or to extract the patterns and rules that drive a community's value-aligned behavior [3].

AI systems, even if they are value-aware, can only be trustworthy if they are capable of explaining the *reasons* for their decisions, so they can be validated and/or audited. The DARPA Explainable Artificial Intelligence (XAI) [4], for instance, aims at creating AI systems whose learned models and decisions can be understood by end users. This includes seeking methods to increase the interpretability of models, designing effective explanation interfaces, and understanding the psychological requirements of effective explanations. In particular, value-aware systems must be capable of explaining the models and justifying decisions taken in a human-understandable manner, in terms of the values and norms that influenced the reasoning process, among others. To this respect, we can draw upon work on explanation-generation in computational legal reasoning [5].

The present work sets out from the s(LAW) [6] framework initially proposed by Arias et al., which allows for modelling legal rules involving ambiguity, and supports reasoning and inference of conclusions based on them. Moreover, thanks to the goal-directed implementation of the underlying Answer Set Programming (ASP) platform [7], s(LAW) is capable of providing justifications of the resulting conclusions (in natural language). We conjecture that the use of frameworks such as s(LAW) allows for addressing several challenges associated with value-aware systems.

To illustrate our approach, we draw upon the problem of school place allocation in educational institutions supported with public funds. This problem has been present in many countries and for many years [8, 9]. Depending on the value system upheld by a public administration governing a certain territory, different legislations exist, even within the same country. For instance, the Spanish Organic Law on Education[2] regulates, in article 84, the criteria for the admission of students in public centers and private subsidized centers and in its second paragraph indicates adjudication criteria. However, since Spain is a politically decentralized country, its autonomous communities (and their educational administrations) have powers to develop these aspects of basic state legislation. In this work, we will analyse the criteria used in the procedures for awarding school places of centers supported with public funds applied in the Spanish autonomous communities "Comunidad de Madrid" and "Ceuta y Melilla" so as to characterise the underlying value system. Setting out from these real-world cases, which are based on the regulations that are currently in force in two corresponding legislations, we outline their representation in s(LAW) and the types of queries that it supports.

[1] https://thegradient.pub/has-ai-found-a-new-foundation

[2] Organic Law 2/2006, May 3, last modified by Organic Law 3/2020, December 29

1. Modeling value-awareness norms in s(LAW)

The modeling we are proposing, following the same strategy used in [6] to represent vague concepts such as discretion, ambiguity or lack of information, allows us to consider its application in (at least) the following three scenarios:

- Given an allocation criterion, automate the process of awarding places (this use case includes advising parents to select the school where they are most likely to get a place as their first choice). In this scenario, the system could, in case of a tie, make decisions to obtain student distributions that guarantee educational equality. However, the degree of freedom of a system, in this scenario, to improve the alignment with a given value is low.
- Given two or more allocation criteria and the semantic function of the educational equality value, determine which criterion is more aligned, i.e., the application of which system would result in a more equitable distribution. In this scenario, if we consider presence of vague concepts, we may find that an allocation criterion is more aligned under some assumptions, but considering other assumptions it is not the most aligned. As a particular use case, we would have the need for schools to select which complementary criteria are the ones that would result in the most equitable outcome.
- Considering that we only have defined the semantic function of the principle of educational equality, we could automatically generate the most appropriate legislation to guarantee an equitable distribution. In this scenario, we could define a priori a series of normative patterns to facilitate the design of the legislation to be applied As an example consider that we provide an allocation criterion without determining the points to assign, and let the system determine the score to receive in each rating range (and eventually even allow the system to define the rating ranges).

2. Conclusions

In this work, we argue that ASP-based representations together with goal-directed inference are an effective means to introduce value and norm-based reasoning into AI Systems. We analysed two real world cases, based on regulations for school place assignments that are currently in force in different autonomous regions of Spain, and characterised the underlying value system. Furthermore, we provided hints on how these real-world cases can be represented in the s(LAW) legal reasoner, and showed general types of queries that can be answered, such as reasoning about school place admission, determining which set of admission criteria produces results more aligned with the equality value, and assisting with the adaptation of admission criteria in order to improve alignment with the equality value when circumstances change.

In summary, the question we wanted to answer is whether it is possible to “measure” the alignment of different normative systems to the values implicit in the right to education, such as equality, equal opportunities, social cohesion and non-segregation, among others. In addition to answering this question from a legal point of view, in this work we have offered different patterns for modeling norms and values, using s(LAW), so that we are able to automate this measurement.

References

[1] N. Montes, N. Osman, C. Sierra, M. Slavkovik, Value engineering for autonomous agents, CoRR abs/2302.08759 (2023). URL: https://doi.org/10.48550/arXiv.2302.08759. doi:10.48550/arXiv.2302.08759. arXiv:2302.08759.

[2] G. di Tosto, F. Dignum, Simulating social behaviour implementing agents endowed with values and drives, in: F. Giardini, F. Amblard (Eds.), Multi-Agent-Based Simulation XIII - International Workshop, MABS 2012, Valencia, Spain, June 4-8, 2012, Revised Selected Papers, volume 7838 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 1–12. URL: https://doi.org/10.1007/978-3-642-38859-0_1. doi:10.1007/978-3-642-38859-0_1.

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

[4] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI Magazine 40 (2019) 44–58. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850. doi:10.1609/aimag.v40i2.2850.

[5] J. Arias, M. Carro, Z. Chen, G. Gupta, Justifications for Goal-Directed Constraint Answer Set Programming, in: Proceedings 36th International Conference on Logic Programming (Technical Communications), volume 325, EPTCS. Open Publishing Association, 2020, pp. 59–72. doi:10.4204/EPTCS.325.12.

[6] J. Arias, M. Moreno-Rebato, J. A. Rodríguez-García, S. Ossowski, Modeling Administrative Discretion Using Goal-Directed Answer Set Programming, in: Advances in Artificial Intelligence, CAEPIA 20/21, Springer International Publishing: Cham, 2021, pp. 258–267. doi:10.1007/978-3-030-85713-4_25.

[7] J. Arias, M. Carro, E. Salazar, K. Marple, G. Gupta, Constraint Answer Set Programming without Grounding, Theory and Practice of Logic Programming 18 (2018) 337–354. doi:10.1017/S1471068418000285.

[8] M. Gelfond, Strong introspection, in: T. L. Dean, K. R. McKeown (Eds.), Proceedings of the 9th National Conference on Artificial Intelligence, Anaheim, CA, USA, July 14-19, 1991, Volume 1, AAAI Press / The MIT Press, 1991, pp. 386–391. URL: http://www.aaai.org/Library/AAAI/1991/aaai91-060.php.

[9] M. Gelfond, Logic programming and reasoning with incomplete information, Ann. Math. Artif. Intell. 12 (1994) 89–116. URL: https://doi.org/10.1007/BF01530762. doi:10.1007/BF01530762.