Noname manuscript No.
(will be inserted by the editor)

Value-aligned AI: Lessons learnt from value-aligned norm selection

Marc Serramia · Maite Lopez-Sanchez · Juan A. Rodriguez-Aguilar

Received: date / Accepted: date

Abstract One of the key challenges involving ethics in AI is that of value alignment. Value alignment aims at designing AI systems whose behaviours conform to (or align with) the society's moral values and its preferences. In order to illustrate how this process can be addressed, we look into a particular problem: that of selecting value-aligned norms. We use this problem to introduce both the quantitative and qualitative methods we have proposed to solve it. Moreover, we draw some general ideas on how they can be cast to other problems. Finally, we share some of the insights in favour of the qualitative (non-utility-based) method.

Keywords Ethics · Value alignment · Norms

1 Introduction

Moral values are the moral objectives worth striving for [18]. Although societies share common values [23], each society has different preferences over them. These intricacies can be formally captured through a value system [13, 26], an object containing all moral values of the society and the preferences among them. When designing AI with ethics in mind, these values and preferences should be taken into account. Indeed, as outlined by the value alignment problem [21], we should seek to prevent that AI systems act in hostile ways towards humans. Therefore, our goal is to ensure that they act in a way that aligns with human values (the value system of our society). Although the value alignment problem is formulated for any Artificial Intelligence (AI) system (as the European Commission assumes in its legal framework on AI [8]), we can consider value alignment problems in more particular scenarios. Thus, we can

Artificial Intelligence Research Institute (IIIA-CSIC)
University of Barcelona
marcserr@iiia.csic.es, maite_lopez@ub.edu, jar@iiia.csic.es

think of the problems of: (i) designing responsible AI systems [9,3], (ii) studying how agents can learn to choose value-aligned actions while interacting with their environment [19,20], or (iii) enacting norms by a system designer to foster value-aligned agent behaviours [27,25]. The first problem above, also referred as that of designing trustworthy AI [7], has attracted most attention in the AI Ethics research community. Responsible AI is meant to consider design principles such as fairness, reliability, accountability, or transparency when designing (and using) intelligent systems. Many research works focus on one of such design principles. Thus, for instance, there is a plethora of works on fairness in machine learning (see [10] for a survey). However, our research focuses on the third problem above, which we call *value-aligned norm selection*. Hence, the goal of this paper is to explain how we have addressed value-aligned norm selection and to share the lessons learned in our approach, as we think they can be useful for other value alignment problems.

2 Value-aligned norm selection

Norms have long been used as a coordination mechanism in societies [2]. The Normative Multi-Agent Systems literature has studied norms in terms of: norm emergence [24,22,31]; off-line norm synthesis [28,1]; and on-line norm synthesis [16,17]. Nonetheless, behaviour is affected by both norms and values [14], therefore enacting norms by considering their value alignment becomes key. Indeed, as noted by [29], since norms regulate the behaviour of agents in a society, aligning these norms with the moral values of the society will result in a value-aligned agent behaviour. When it comes to norms and value alignment, [15] optimises norm parameters to *synthesise* norms that maximise norm value-alignment. Our approach in [27,25] takes a system designer perspective and focuses on how to select the norms to enact that best align with the values of the society. We call this problem the value-aligned norm selection problem.

When considering value-aligned norm selection, two objects are important, namely, the norms and the moral values. Although the literature in normative Multi-Agent Systems has proposed alternative norm definitions (see for example [12,6]), norms in this context are typically defined to regulate agents' actions through the specification of deontic operators (i.e., prohibition, permission, and obligation) over those actions. For instance, in a border control at an international airport, we may impose a norm n_1: "Obliging all passengers to show their passport" or an alternative norm n_2: "Permitting them to cross the border". We refer to the set of norms to enact as a *norm system*. Subsequently, to guide norm system selection, we take into account the value alignment of the norms. We do so by considering the values each norm promotes/demotes and the preferences among these values. Following previous example, n_1 supports a "free movement" moral value (v_1), whereas n_2 supports a "safety" (v_2) moral value. Then, deppending on the preferences over these values (that is, if "free movement" is preferred over "safety" or the other way arround), the norms to enact should be chosen accordingly.

Note though that norms can be interrelated: they can be mutually exclusive (e.g., a norm prohibiting an action is exclusive with a norm obliging it) or they can be redundant (e.g., a norm may generalise another one by having a broader scope as, in our airport example, we may also consider norms requiring the passport to some passengers). Thus, not all possible norm systems constitute a feasible solution as we want to avoid exclusive and redundant norms inside a norm system. We refer to a *sound* norm system as that free of exclusive and redundant norms. In this manner, the value-aligned norm selection problem consists on selecting the most value-aligned sound norm system.

At this point it may be worth noticing that this problem specification assumes that preferences over the moral values will be provided, and that these value preferences will guide the selection of the norms to enact. Although moral theories (such as Kantian [11], Rawlsian [4], or Lockean [30]) have long discussed how norms should be established in a society, the rationale behind our assumption is that a society should be regulated according to its moral values, so that individuals in this society are instructed to behave in accordance to them.

3 Utilitarian value-alignment

A possible way to tackle value-aligned norm selection is by using utilities, which we addressed in [27]. In this approach, the decision maker is required to provide norm-value promotion/demotion degrees linking each norm to each value. Positive degrees mean that the norm promotes the value while negative ones mean that it demotes it. Then, these pairwise norm-value promotion/demotion degrees are combined to assess the overall value alignment utility of each norm, and to subsequently find the sound norm system with greater overall utility. In more detail, this is done in three stages. First, utilities are assigned to values considering the value preferences in the value system: the more preferred a value in the value system, the greater its utility. Afterwards, norm utilities are computed by combining the value utilities and the norm-value promotion/demotion degrees (those provided by the decision maker relating each norm to each value). Norm utility is increased with value promotion while it is decreased with value demotion. Having all norm utilities, the utility of a norm system results from aggregating the utilities of its norms. Hence, it is possible to find the sound norm system with maximum utility using optimisation techniques.

Nonetheless, utilitarian value-aligned norm selection is far from being perfect. On one hand, the decision maker is required to provide input regarding the relation between norms and values in the form of promotion/demotion degrees. Finding these degrees is difficult and can lead to biases in the selection. On the other hand, the additivity of utilities can lead to undesirable outcomes. This becomes more apparent when we consider mediocre norms, that is, norms that neither promote nor demote the most preferred values but still have small but positive utility because they promote other values. If

several of these mediocre norms are considered, their overall cumulative utility can surpass that of a useful norm promoting the most preferred values. This means that a norm system full of mediocre norms would be selected instead of another norm system containing less norms but useful ones. These shortcomings are independent of the framework of the problem (i.e., they are independent of the specific norms and values actually used). Instead, they are direct consequences of the nature of utilities (and their additivity). Thus, to overcome them, we proposed to solve the problem by following the qualitative approach introduced next.

4 Qualitative value-alignment

To avoid the problems of the utilitarian approach, we argue that a qualitative approach is more convenient, such as the one in [25]. Here, the decision maker is required to provide norm-value relations in qualitative terms (such as labels) instead of numerical degrees. This task is simpler for the decision maker as labels allow to have different degrees of expressiveness[1]. Once we know the relation between norms and values, we basically work with these relations and the value preferences without translating them into numbers. Overall, the process is divided into two steps. First, knowing both how norms relate to values and the preferences among these values, we transform value preferences into individual norm preferences. This can be achieved by exploiting or adapting social choice preference functions, for example, in [25] we adapt lex-cel [5]. The principle we adhere to is two-fold: (i) the more preferred the values the norm promotes, the more preferred the norm; whereas (ii) the more preferred the values the norm demotes, the less preferred the norm. This procedure also ensures that the obtained norm preferences embody value alignment, meaning that the more preferred a norm, the more value-aligned. The next step consists in transforming individual norm preferences into preferences over sets of norms (i.e., norm systems). Similarly, a qualitative operator based on lexicographical order can be applied to translate the principle of: the more preferred the norms in a norm system, the more preferred the norm system. Since norm preferences embody value alignment, we conclude that the more preferred a norm system, the more value aligned its norms. Hence, finding the solution to the value-aligned norm selection problem amounts to building norm system preferences and finding the most preferred norm system that is sound.

5 Discussion

Applying norm-value alignment may raise some questions about the level of abstraction that the system designer should consider. In this regard, we assume

[1] It may be worth noticing that although [25] only considers promotion and no promotion, we can conveniently increase its expressivity by considering various degrees of promotion and demotion.

the system designer enacts the role of the policy maker and has enough domain knowledge to be able to define: the set of candidate norms to choose from; how norms relate to other norms; if they promote or demote moral values; as well as the preferences among these values. Therefore, each candidate norm should be valid (acceptable) when considered individually, since the problem we tackle here is that of finding the (sound) subset of norms that is most aligned to the values at hand. However, this does not mean that the policy maker may specify norms at different levels of abstraction, since one of the relationships between norms that our methods handle is, precisely, the generalisation. Our methods will simply choose those norms whose level of abstraction best promotes the most preferred moral values.

Additionally, note that our quantitative method in Section 3 should not be seen as a tight implementation of Consequentialist ethical theories where the consequences of actions are pondered. Instead, our methodderives norm utilities from their support to moral values and the relative preferences over those values.

Finally, it may also be worth discussing that, even though we differentiate our contribution from the research focusing on the paradigmatic values of trustworthy AI (i.e. fairness, transparency, etc.), our approach is general enough to include them. Hence, we can help a policy maker choose norm systems that align with fairness and transparency, if these are the values the policy maker intends to bring about. Furthermore, if the values that are considered pose some conflicts, then, the value preferences in the value system at hand will guide norm selection. Thus, for example, when considering transparency and security, there might be norms that promote one of them but demote the other one (e.g., a norm obliging a senior official to report their meetings promotes transparency but demotes security). Nonetheless, our norm selection approach is expected to choose those norms that promote the most preferred value.

6 Conclusions

This paper discusses our research revolving around the selection of those norms that are most aligned with moral values. Specifically, we introduce two distinct methods –a quantitative and a qualitative method– for norm selection. Although our work is centred on norm selection, the conclusions and experiences we formulate can be useful for other value-alignment problems since, in general, we can reformulate the problem in terms of any set of candidate elements, a value system, and how those elements relate to the moral values. Then, any problem requiring to select value-aligned elements can be solved by employing any of the two proposed methods. However, through our research we have observed that quantitative (utilitarian) approaches to value alignment can become problematic. Firstly, assigning utilities can turn out to be difficult and arbitrary. Secondly, it may be necessary to face the risk of selecting mediocre elements due to utility additivity. While we have only detected these

problems in norm selection, they are inherent properties of utilities and independent of norms. Hence, it is reasonable to think that they will emerge in other value-alignment problems. To avoid these problems we advocate for qualitative approaches. In particular, the qualitative approach is fairly easy to adapt no matter the value preferences considered or the type of link between the elements and the values (be it a more expressive link using labels or just a binary related/not related). As long as the input is qualitative, we can apply the same general resolution. That is, transforming value preferences into element preferences, to subsequently transform element preferences into subset preferences. Then, we can select the most value aligned elements from these preferences. In both cases, if not all solutions are feasible, we can then constrict the selection to only feasible sets.

References

1. Ågotnes, T., Van Der Hoek, W., Sierra, C., Wooldridge, M.: On the logic of normative systems. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI '07, pp. 1175–1180 (2007)
2. Azar, O.H.: What sustains social norms and how they evolve?: The case of tipping. Journal of Economic Behavior & Organization **54**(1), 49–64 (2004)
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020). DOI https://doi.org/10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103
4. Beitz, C.R.: Rawls's law of peoples. Ethics **110**(4), 669–696 (2000)
5. Bernardi, G., Lucchetti, R., Moretti, S.: Ranking objects from a preference relation over their subsets. Social Choice and Welfare **52**(4), 589–606 (2019). DOI 10.1007/s00355-018-1161-1. URL https://doi.org/10.1007/s00355-018-1161-1
6. Boella, G., van der Torre, L.: Regulative and constitutive norms in normative multiagent systems. Proceedings of KR'04 pp. 255–265 (2004)
7. Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., Yeung, K.: Trustworthy ai. In: Reflections on Artificial Intelligence for Humanity, pp. 13–39. Springer (2021)
8. European Comission: Legal framework on AI. https://digital-strategy.ec.europa.eu/en/library/propo sal-regulation-laying-down-harmonised-rules-artificial-intelligence. Accessed: 2021-06-11
9. Dignum, V.: Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature (2019)
10. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, pp. 329–338 (2019)
11. Johnson, R., Cureton, A.: Kant's moral philosophy. In: E.N. Zalta (ed.) The Stanford encyclopedia of philosophy, spring 2018 edn. (2018). URL https://plato.stanford.edu/entries/kant-moral/
12. López y López, F., Luck, M., d'Inverno, M.: Constraining autonomy through norms. In: AAMAS, pp. 674–681. ACM (2002)
13. Luo, J., Meyer, J.J., Knobbout, M.: Reasoning about opportunistic propensity in multi-agent systems. In: AAMAS 2017 Workshops, Best Papers., pp. 1–16 (2017)
14. Mercuur, R., Dignum, V., Jonker, C., et al.: The value of values and norms in social simulation. Journal Artificial Societies and Social Simulation **22**(1), 1–9 (2019)

15. Montes, n., Sierra, C.: Value-guided synthesis of parametric normative systems. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2021), pp. 907–915. International Foundation for Autonomous Agents and Multiagent Systems (2021)
16. Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Vasconcelos, W., Wooldridge, M.: On-line automated synthesis of compact normative systems. ACM Transactions on Autonomous and Adaptive Systems (TAAS) **10**(1), 2:1–2:33 (2015)
17. Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Wooldridge, M., Vasconcelos, W.: Automated synthesis of normative systems. In: AAMAS 2013, pp. 483–490 (2013)
18. van de Poel, I., Royakkers, L.: Ethics, Technology, and Engineering: An Introduction. Wiley-Blackwell (2011)
19. Rodriguez-Soto, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: A structural solution to sequential moral dilemmas. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1152–1160 (2020)
20. Rodriguez-Soto, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: Multi-objective reinforcement learning for designing ethical environments. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence, pp. 1–7 (2021)
21. Russell, S.: Human compatible: Artificial intelligence and the problem of control. Penguin (2019)
22. Savarimuthu, B.T.R., Purvis, M., Cranefield, S., Purvis, M.: Mechanisms for norm emergence in multiagent societies. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, AAMAS '07, pp. 173:1–173:3. ACM, New York, NY, USA (2007). URL http://doi.acm.org/10.1145/1329125.1329335
23. Schwartz, S.: An overview basic human values: Theory, methods, and applications introduction to the values theory. Jerusalem Hebrew University (2006)
24. Sen, S., Airiau, S.: Emergence of norms through social learning. In: IJCAI, pp. 1507–1512 (2007)
25. Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: A qualitative approach to composing value-aligned norm systems. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1233–1241 (2020)
26. Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A., Morales, J., Wooldridge, M., Ansotegui, C.: Exploiting moral values to choose the right norms. In: Proceedings of the 1st Conference on artificial intelligence, ethics and society (AIES'18), pp. 1–7 (2018). DOI 10.1145/3278721.3278735
27. Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Rodriguez, M., Wooldridge, M., Morales, J., Ansotegui, C.: Moral values in norm decision making. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18), pp. 1294–1302. International Foundation for Autonomous Agents and Multiagent Systems (2018)
28. Shoham, Y., Tennenholtz, M.: On social laws for artificial agent societies: off-line design. Artificial Intelligence **73**(1-2), 231–252 (1995)
29. Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., Perello-Moragues, A.: Value alignment: A formal approach. In: Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS 2019 (2019)
30. Simmons, A.J.: The Lockean theory of rights. Princeton University Press (2020)
31. Sugawara, T.: Emergence and stability of social conventions in conflict situations. In: T. Walsh (ed.) IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011, pp. 371–378. IJCAI/AAAI (2011). DOI 10.5591/978-1-57735-516-8/IJCAI11-071. URL https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-071