Towards Automating the Design of Value-Aligned Clinical Protocols

Extended Abstract

Manel Rodriguez-Soto
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain
manel.rodriguez@iiia.csic.es

Nardine Osman
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain
nardine@iiia.csic.es

Carles Sierra
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain
sierra@iiia.csic.es

Rocio Cintas-Garcia
Hospital del Mar Research Institute (IMIM)
Barcelona, Spain
rncintasgarcia@psmar.cat

Cristina Farriols-Danes
Hospital del Mar Research Institute (IMIM)
Barcelona, Spain
cfarriols@psmar.cat

Montserrat Garcia-Retortillo
Hospital del Mar Research Institute (IMIM)
Barcelona, Spain
mgarciaretortillo@psmar.cat

Silvia Minguez-Maso
Hospital del Mar Research Institute (IMIM)
Barcelona, Spain
sminguez@psmar.cat

Jordi Martinez-Roldan
Hospital Sant Joan de Deu (SJD)
Barcelona, Spain
jordi.martinezr@sjd.es

ABSTRACT

Clinical protocols are of great use in all medical fields, but their design and evaluation is complex and time-consuming. One of their major complexities is that they need to respect all bioethical values. For those reasons, in this paper, we address the problem of automating the design of clinical protocols that are in alignment with bioethical values. Following the AI alignment literature, we propose an algorithm to design value-aligned protocols in several steps: *BPR (Bioethical Protocol Recommender)*. First, BPR learns the implicit bioethical value definitions of clinicians. Thereafter, BPR can apply these definitions to evaluate and compare potential actions for any given patient. With that, BPR builds a protocol that recommend those actions with maximum alignment with respect to all bioethical values, applying multi-objective optimisation techniques.

KEYWORDS

value awareness; value alignment; clinical protocols; medical protocols

ACM Reference Format:
Manel Rodriguez-Soto, Nardine Osman, Carles Sierra, Rocio Cintas-Garcia, Cristina Farriols-Danes, Montserrat Garcia-Retortillo, Silvia Minguez-Maso, and Jordi Martinez-Roldan. 2025. Towards Automating the Design of Value-Aligned Clinical Protocols: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Values have been extensively studied across various fields, including psychology, sociology, and philosophy, as they are identified as key motivators that influence human behaviour and social interactions [5, 10, 16]. With the rapid integration of AI into every aspect of our lives, a major challenge is developing AI systems whose behaviour, or the behaviour they facilitate, aligns with human values. This is known as the value-alignment problem [3, 4, 7]. Given that norms have traditionally been used in multiagent systems as means for mediating behaviour, a number of methodologies have been proposed for assessing the alignment of norms and our values [8, 9, 12, 13, 18]. Clinical protocols can be understood as one example of such norms. Clinical protocols are structured guidelines that provide healthcare professionals with standardised procedures for treating patients. They are usually designed to ensure that the four major bioethical values are respected: beneficence (ensuring the patient's benefit), non-maleficence (minimising the patient's harm), justice (treating all patients equally), and autonomy (respecting the patient's considerations) [2]. Despite the existing work on value-alignment of norms [8, 17, 18], there is still a gap in the design of such norms for optimal alignment, especially considering potential value conflicts. In the case of clinical protocols, value conflicts often arise amongst these four basic bioethical values, such as conflicts between ensuring beneficence and respecting the patient's wishes. Against this background, we tackle the challenge of designing clinical protocols that consider all bioethical values by means of the following contributions:

- We present a novel mechanism for learning value alignment semantics of clinicians, BAAL (Fig. 1 (left)).
- Building on BAAL, we design the novel algorithm BPR (Fig. 1) that recommends value-aligned actions and protocols.

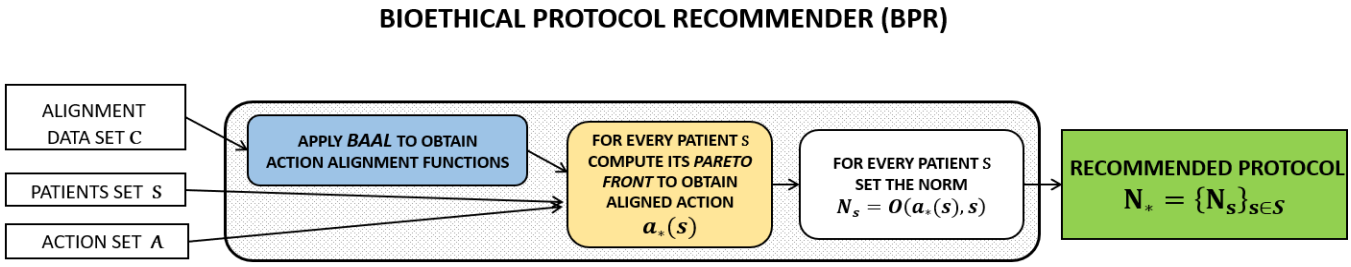

Figure 1: Our whole algorithm for recommending protocols that are value alignment with all four bioethical values. Rectangles stand for objects while rounded rectangles correspond to processes.

2 BACKGROUND

An agent is considered to act in alignment with a given value if their behaviour promotes this particular human value. In this work, we adopt the definitions of [11, 19], interpreting values as preferences that enable the comparison of different possible states of the world:

Definition 2.1 (Action alignment). Let $\mathcal{S}$ be a set of states, and $\mathcal{A}$ a set of actions. The action alignment function is defined as a function $align : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{V} \rightarrow [-1, 1]$ where $\mathcal{V}$ is a set of values. We want the range of alignment to be $[-1, 1]$ so that positive alignment represents the action promoting the value, negative alignment represents the action demoting the value, and an alignment of zero represents the action not affecting the value.

3 PROTOCOLS RECOMMENDER ALGORITHM

This section is devoted to explain the three steps of our algorithm for providing value-aligned bioethical protocols.

3.1 Estimating Action Alignment

The first step of our algorithm, BAAL (see Figure 1 left), learns the action alignment functions $align(s, a, s', V)$ that define what it mean for each value V to be respected in each given case $\langle s, a, s' \rangle$. For that, we require a corpus C_v of labelled alignment data in the form of state-action-state-alignment tuples:

$$C_v = \{s_i, a_i, s'_i, align(s_i, a_i, s'_i, v)\}_i, \tag{1}$$

for each bioethical value v. From each C_v, we apply a regression modelling algorithm (dependent on the corpus) to learn the following conditional expectations of action alignment:

$$\widehat{SAalign}_v(s, a) = \mathbb{E}[align(S, A, S', v) \mid S = s, A = a]. \tag{2}$$

With our expected alignment functions computed, we can proceed to the following steps of our algorithm for recommending aligned actions and protocols.

3.2 Recommending Actions

Next, making use of our estimated state action alignment functions $\widehat{SAalign}$, we expect our algorithm to recommend actions that align with all values as much as possible, without explicitly prioritising any value. This approach is in accordance with principialism [1], the philosophical theory originating the four bioethical values. With principialism into account, we consider that the most adequate option is to compute the *Pareto front* (PF) of optimal actions with respect to each bioethical value. Then, we let the user select from the PF, a standard practice in the multi-objective literature [6, 14, 15].

Thus, given a set of patient states $\mathcal{S}$, and a set of candidate actions $\mathcal{A}$ (both input in Figure 1) our algorithm computes the Pareto front $PF(s)$ of each patient for the set of state-action alignment vectors

$$\mathcal{SVA} \doteq \{\widehat{SAalign}_v(s, a)\}_{v,a}, \tag{3}$$

Thereafter, the user can select whichever action $a_* \in PF(s)$ they prefer, since all of them are Pareto-optimal. This final action a_* is the one recommended by our algorithm for the given patient s.

3.3 Recommending Protocols

Finally, to design a value-aligned protocol, for every state s, we need to permit only those actions that achieve the maximum amount of alignment. That is, the Pareto-optimal actions that the previous step (Figure 1 (middle)) finds. As a simplified approach to compute a value-aligned protocol, for each possible patient state s, we set the norm

$$N_s \doteq O(a_*(s), s), \tag{4}$$

where $a_*(s) \in PF(s)$ belongs to the Pareto front of state s, and $O(a, s)$ indicates that action a is *obligatory* under state s. Thus, our algorithm returns the recommended protocol $N_* = \{N_s\}_{s \in \mathcal{S}}$ such that for every state $s \in \mathcal{S}$ it recommends norm N_s. This protocol, by construction, will be value-aligned to all bioethical values.

4 CONCLUSIONS

This paper tackles the open problem of value-aware decision-making and value-aware protocol design in the medical field. Our novel contributions are in two fronts. First, based on the literature, we have formalised a method (BAAL) for learning the value alignment semantics of bioethical values. Second, we have provided an algorithm (BPR) for designing value-aligned protocols. Our algorithm allows us to compute the expected value alignment of any medical action. Therefore, our algorithm can discard any action that would never be chosen due to its low degree of alignment, and recommend to the user only actions that are Pareto-optimal with respect to all four bioethical values. In future work, we expect to analyse the effects of BPR on protocols from Hospital del Mar, Barcelona.

ACKNOWLEDGMENTS

This work has been supported by the EU-funded VALAWAI (# 101070930) project and the Spanish-funded VAE (# TED2021-131295B-C31).

REFERENCES

[1] T.L. Beauchamp and J.F. Childress. 1979. *Principles of Biomedical Ethics.* Oxford University Press. https://books.google.es/books?id=_CujQgAACAAJ

[2] Tom L. Beauchamp and James F. Childress. 1979. *Principles of biomedical ethics / Tom L. Beauchamp, James F. Childress.* Oxford University Press New York.

[3] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. 2021. Trustworthy AI. In *Reflections on Artificial Intelligence for Humanity*. Springer, 13–39.

[4] European Comission. 2021. Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206. Accessed: 2021-06-29.

[5] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (09 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[6] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2022. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *Autonomous Agents and Multi-Agent Systems* 36 (2022).

[7] IEEE. 2019. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html. Accessed: 2021-06-29.

[8] Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé (Eds.). ACM, 907–915. https://doi.org/10.5555/3463952.3464060

[9] Nieves Montes and Carles Sierra. 2022. Synthesis and Properties of Optimally Value-Aligned Normative Systems. *J. Artif. Intell. Res.* 74 (2022), 1739–1774. https://doi.org/10.1613/jair.1.13487

[10] Laura Parks-Leduc and Russell Guay. 2009. Pesonality, values, and motivation. *Personality and Individual Differences* 47 (11 2009), 675–684. https://doi.org/10.1016/j.paid.2009.06.002

[11] Manel Rodriguez-Soto, Nardine Osman, Carles Sierra, Paula Sánchez Veja, Rocio Cintas Garcia, Cristina Farriols Danes, Montserrat Garcia Retortillo, and Silvia Minguez Maso. 2024. User Study Design for Identifying the Semantics of Bioethical Principles. In *Second International Workshop on Value Engineering in Artificial Intelligence (VALE) at the European Conference on Artificial Intelligence.*

[12] Manel Rodriguez-Soto, Marc Serramia, Maite López-Sánchez, Juan A. Rodriguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. 2023. Encoding Ethics to Compute Value-Aligned Norms. *Minds and Machines* (11 2023). https://doi.org/10.1007/s11023-023-09649-7

[13] Manel Rodriguez-Soto, Marc Serramia, Maite López-Sánchez, and Juan Rodríguez-Aguilar. 2022. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology* 24 (03 2022). https://doi.org/10.1007/s10676-022-09635-0

[14] Diederik Roijers and Shimon Whiteson. 2017. *Multi-Objective Decision Making.* Morgan and Claypool, California, USA. http://www.morganclaypool.com/doi/abs/10.2200/S00765ED1V01Y201704AIM034 doi:10.2200/S00765ED1V01Y201704AIM034.

[15] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2019. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34, 1 (Dec. 2019), 52. https://doi.org/10.1007/s10458-019-09433-x

[16] Shalom Schwartz. 2006. An Overview Basic Human Values: Theory, Methods, and Applications Introduction to the Values Theory. *Jerusalem Hebrew University* (2006).

[17] Marc Serramia, Maite López-Sánchez, and Juan A. Rodríguez-Aguilar. 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, Auckland, New Zealand, May 9-13, 2020*, Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, 1233–1241.

[18] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. 2021. Value alignment: a formal approach. *CoRR* abs/2110.09240 (2021). arXiv:2110.09240 https://arxiv.org/abs/2110.09240

[19] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perello-Moragues. 2019. Value alignment: A formal approach. *Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS 2019* (2019).