Towards Value Awareness in the Medical Field

Manel Rodriguez-Soto[1], Nardine Osman[1], Carles Sierra[1], Paula Sánchez Veja[2], Rocio Cintas Garcia[2], Cristina Farriols Danes[2], Montserrat Garcia Retortillo[2] and Silvia Minguez Maso[2]

[1] *Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain*

[2] *Hospital del Mar Research Institute (IMIM), Barcelona, Spain*

Keywords: Value Awareness, Value Alignment, Medical Protocols, Medical Corpus.

Abstract: This position paper aims to illustrate how models and mechanisms can be designed to support value-aware decision-making in the medical field. Such models and mechanisms allow for assessing the alignment of specific behaviours with human values, which could help medical personnel decide when to follow or break a protocol and help relevant boards decide when and how to update existing protocols. AI supporting decision-making in medicine is not new. Yet, AI that raises awareness about the alignment of medical decisions concerning human values is novel despite the vital importance of bioethics in the field. This paper presents a proposal for the formalisation of values and the design of models and mechanisms that raise value awareness in the medical field.

1 INTRODUCTION

With the growing risks and fears of AI, developing ethical AI has become a top objective of many governmental bodies, organisations, and AI scientists. One approach for achieving ethical AI is ensuring AI that aligns with human values. Stuart Russell argues that we should change the overarching goal of AI from "intelligence" to "intelligence provably aligned with human values" (Russell, 2019), a topic now known as the *value alignment problem*.

The ultimate goal of this research is to develop models and mechanisms for value-aware situation analysis and decision-making. The objective is to have AI systems that are aware of our value systems and can explain their behaviour or understand the behaviour of others in terms of those value systems. In other words, in addition to AI being aware of human values and reasoning with them, it can help humans become aware of the alignment of their behaviour with different values. The latter is what this paper is working towards, using AI to support medical decision-making by raising value awareness.

The Hippocratic Oath is still considered today to be a cornerstone and foundation of the medical profession across the world. It reflects the moral values that define the medical profession, and some (Askitopoulou and Vgontzas, 2018) considered it to have exemplified some of the fundamental modern ethical principles (such as beneficence, non-maleficence and confidentiality) that have developed since 1970 and have been incorporated into the undergraduate and postgraduate medical curriculum, residency training, and continuous professional education across the Anglo-European world (Ngan and Sim, 2021). However, despite the extensive AI tools that support medical decision-making today, there is a complete lack of tools that analyse decisions from the perspective of their alignment with human values.

The objective of this position paper is to develop AI that is capable of explaining the alignment of certain medical decisions with values. Such an AI will help medical personnel decide when to follow or break a protocol and help relevant boards decide when and how to update existing protocols.

Given the current background in developing value-aligned AI (Sierra et al., 2021; Montes and Sierra, 2021; Rodriguez-Soto et al., 2022; Montes and Sierra, 2022; Rodriguez-Soto et al., 2023), we set out to develop tools that are application-driven, allowing us to address real-life problems. This paper opens with Section 2 by presenting the chosen medical protocol of our use case, followed by Section 3 which introduces the construction of the corpus that our AI tools will build upon. Section 4 then presents our plan for the formal specification of values and the

Rodriguez-Soto, M., Osman, N., Sierra, C., Sánchez Veja, P., Cintas Garcia, R., Farriols Danes, C., Garcia Retortillo, M. and Minguez Maso, S.
Towards Value Awareness in the Medical Field.
DOI: 10.5220/0012588600003636
In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 1391-1398
ISBN: 978-989-758-680-4; ISSN: 2184-433X

development of a moral reasoner, before concluding with Section 5.

2 THE MEDICAL PROTOCOL

2.1 Medical Protocols

Clinical guidelines are an agreed framework outlining the care that will be provided to patients in a designated area of practice. These documents provide information and recommendations about therapeutic interventions, specify procedures to be followed in defined situations, and are based on an assessment of the current best evidence of clinical and cost-effectiveness. Their purpose is to support a clinician in the management of a specific clinical problem but also they can assist patients in making informed decisions and enhance the communication between the patient and the healthcare professional.

Medical protocols consist of a written set of instructions that describes the approved and recommended steps of a particular act or sequence of clinical events. They are more explicit and specific in their detail than guidelines because they specify who does ‘what’, ‘when’ and ‘how’ once a clinical management decision has been made.

The medical protocols of Hospital del Mar are documents that are elaborated by the professionals involved in the healthcare process and validated by the head of the service. The hospital has more than 1,600 healthcare protocols, of which more than 800 are medical protocols and more than 550 are nursing protocols. To identify a wide number of medical protocols in which ethical conflicts may appear, some examples were analysed and we decided to focus on Hospital del Mar’s Therapy Intensity Level Scale, which we describe next, as an example to identify possible bioethical dilemmas.

2.2 The NIT Protocol

Between 2019–2020, a working group led by the Mortality Commission of Hospital del Mar initiated a project to adapt guidelines from the geriatric field that could provide support in those cases where there was a general ethical conflict between patients, family members and healthcare professionals. The working group resulted in designing a Therapy Intensity Level scale (NIT scale), a classification system that was adapted from the Rogers Memorial Veterans Hospital and consists of 5 levels. This dynamic tool allows professionals to update information throughout patient care and helps resolve therapeutic decision aspects quickly. The classification is based on a shared decision-making process agreed upon with the healthcare professionals and the patients (or their representatives) to guarantee adequate coherence between the patient’s wishes, values and preferences and treatment intensity. The NIT scale is used to identify which actions are susceptible to be adopted depending on the therapeutic objective (prolong patient’s life expectancy, enhance patient’s comfort or increase their independence, ...).

The NIT level categorises treatments into 5 levels:

- **NIT 1:** This is the no-limitation treatments level. All measures and efforts that prolong a patient’s survival are accepted.
- **NIT 2:** This is the intensive treatment. The long-term survival is the main objective.
  - **NIT 2A:** This accepts all measures except CPR.
  - **NIT 2B:** This does not accept CPR but contemplates semi-criticals unit and some intensive therapies.
- **NIT 3:** This is an intermediate level that accepts complementary examinations and non-invasive treatments. CPR, ICU and semi-criticals unit are excluded.
- **NIT 4:** This is the conservative treatment: symptom control and palliative care. Only symptomatic and empirical treatments are included. Complementary explorations are prevented. CPR, ICU or Semi-criticals unit must be avoided.
- **NIT 5:** This is applied to end-of-life patients situation. Comfort care is the main purpose. Only comfort measures and treatments focused on providing comfort, quality of life and dignity (instead of extending life) must be taken.

2.2.1 Value Awareness and the NIT Protocol

As we saw above, each NIT level has a number of norms built within it, like giving a recommendation for or against some actions (such as moving the patient to an ICU unit, or applying CPR). As such, an AI system can check whether potential actions follow or not the NIT protocol. However, more interestingly, we plan to develop an AI system that can take into consideration important values (such as those that we present shortly in Table 3) and evaluate each potential action with respect to those values. We envision the medical professional to provide the AI with a set of potential actions that they are considering to perform. These are selected from a predefined list of actions (such as those presented in Table 2). The system can then provide information on the alignment of each action with the NIT protocol along with its alignment with important values. Some of the important values, which we discuss shortly in Section 3, are the basic values of autonomy, beneficence, non-maleficence and justice, along with additional values that are deemed important to the hospital (e.g. cost efficiency) or the patient (e.g. no pain). In other words, an AI can raise value awareness for the medical professional's decision-making process by analysing which values would a given potential action promote or demote, and to what degree.

Furthermore, as we illustrate in Section 4, we can deduce the alignment of medical protocols from the alignment of actions. As such, an AI system will also be capable of providing feedback on the alignment of medical protocols, like the NIT protocol, with certain values. This helps raise value awareness at the management level to help with decisions on when and how should medical protocols evolve.

3 BUILDING THE CORPUS

One of the main challenges faced in this medical use case is that the necessary data needed for reasoning about values is not currently available. It either exists in a non-digitised form or as part of the medical personnels' know-how. To this end, the first step was to work on building a corpus that can later on be used (as we show in Section 4) for reasoning about values.

We commenced this line of work by compiling a few entries of medical cases, with each entry consisting of four main parts:

1. The criteria that describe a patient's medical state, such as their age and pathology, as described in Table 1. We note that we decided to focus on general criteria that could be deduced from the medical files and that would help with value-based analysis and decision-making.
2. The actions available for the medical personnel to choose from, indicating which action was performed or not, whether it was effective or not, and whether it is aligned with the NIT protocol or not. The set of actions is pre-defined, and it is presented in Table 2.
3. The criteria that change as a result of taking an action, such as the change in expected survival, if any. Those are marked in Table 1 under the 'Changes with actions' column.
4. The relevant values that the AI reasoner must consider in this medical use case, and whether each action promotes, demotes, or does not affect that value. The selected values are the basic four bioethical values that the medical personnel are trained to respect (Beauchamp and Childress, 1979), plus additional values, such as values that are deemed important to the hospital (e.g. cost efficiency) or values that are deemed important to the patient (e.g. suffer no pain, better quality of life). All of these values, presented in Table 3, have been provided by the medical doctors of the NIT unit. We acknowledge an overlap between basic values and patient values, which requires further analysis.

In addition to the above four main parts, each case has an anonymised case number, a date, and the assigned NIT level.

As an example, we provide the details of one entry from our corpus, that of Case #4. Case #4 entered the hospital under NIT level 4. The patient was 73 years old, with complex chronic disease, short-term survival with an expected survival of less than 6 months, high Frail-VIG, without social support, with a slight functional independence (with a Barthel Index of 91–99), and a slight-moderate cognitive deterioration. The action 'Mild: TC/Transf/Picc/Enteral Nutrition/others' (in this case, it was a catheter peritoneal insertion) was considered not useful but was taken by the medical doctor to improve patient comfort. According to the analysis of the medical doctors populating the corpus, the result of taking this action promoted the values 'autonomy' and 'beneficence', but demoted the values 'non-malificience' and 'justice'. It also demoted the value 'cost-efficiency', but promoted the values 'symptoms controlled (no pain)' and 'better quality of life'. We note that the action was not aligned with the assigned NIT level, which should have been NIT 3.

4 REASONING ABOUT VALUES IN THE MEDICAL USE CASE

In what follows, we present how models and mechanisms can be developed to make use of the above corpus for reasoning about values. The first step will consist of representing biomedical values as formal objects. Such a formalisation will allow us to have a transparent, precise and computational definition of values to determine which behaviours are aligned with them. Then, as a second step, we foresee two different approaches that will be explored further in future work to determine value alignment. The first uses machine learning (Murphy, 2022; Jordan and Mitchell, 2015) to learn from the compiled corpus data, while the second uses symbolic reasoning (McCarthy, 1999; Montes and Sierra, 2021; Curto et al.,

Table 1: Patient criteria.

Criteria	**Description**	**Changes with actions**
Criterion 1: Age (S)	Patient's age	
Criterion 2: Patient with Complex Chronic Diseases (CCD)	Measures if the patient has one or more chronic diseases with at least one being permanent, leaving lingering disability, being non-reversible, or coexisting with a psychological illness.	
Criterion 3a: Short-term survival	Measures if the patient has an advanced chronic disease with a expected survival rate of less than 12-18 months that requires palliative care.	✓
Criterion 3b: Expected survival	It is an estimation, in months, of the expected survival of the patient.	
Criterion 4: Frail-VIG [Scale: Spict]	Fragility Index validated in the geriatric population. Based on different variables, it offers a frailty evaluation tool for rapid assessment. Frail-VIG establishes the degree of frailty of the patient. This criterion has a reliable mortality predictive capacity.	✓
Criterion 5: Clinical Risk Groups (CRG)	A categorical classification system that uses administrative data to identify patients with chronic health conditions. Stratifies the population according to morbidity groups in four different levels: (0) Promotion & Prevention of diseases. (1) Self-management support: between 70–80% of patients are able to look after their own health efficiently and conveniently (selfcare). (2) Illness management: high risk patients that need illness management coordinated with the professionals. (3) Case management: the patient needs a case manager to coordinate the medical care.	
Criterion 6: Social support (NS)	It considers if the patient has social support (family, friends) to offer support functions (emotional, instrumental, ...)	
Criterion 7: Functional independence (Barthel Index) [Scale: Berthel]	The Barthel Index is a 10-item instrument used in the evaluation of functional independence in personal activities of daily living (ADL). It measures the capacity of a person for the execution of basic activities in daily life (feeding, bathing, ambulation, bladder and bowel control, ...).	✓
Criterion 8: Patient's advanced directives (written or oral)	For patients who have decision-making capacity. Referred if there is a signed document or the patient has mentioned their desires regarding treatment decisions. It includes when the patients identify whom they want to make decisions on their behalf when they cannot do so themselves.	
Criterion 10: Cognitive deterioration	When the patient suffers cognitive impairment (confusion, memory loss, difficulty understanding or speaking, problems with concentration…)	
Criteria 11: Comfort	A dynamic state characterised by absence of pain, emotional and physical distress and symptom control.	✓

2022) to reason about actions and protocols and their alignment with values.

4.1 Value Representation

Regardless of which approach we follow for reasoning about values, we will need to have a formal representation of values to enable this computational reasoning. We commence with the four biomedical values of Beauchamp and Childress' principialism (Beauchamp and Childress, 1979) (Values 1–4 in Table 3). As agreed upon by the biomedical community, at least these four values provide the best framework for ethical analysis in biomedical scenarios (Veatch, 2020).

Our first step towards formalising the four biomedical values is first to categorise them following the proposed outline by Veatch in (Veatch, 2020). Veatch states that biomedical ethics' four main values can be divided into two categories: *consequence-based* values and *duty-based* values. To behave in alignment with a given value has a separate definition for each category, as we show next.

- **Consequence-based values:** An action is aligned with a consequence-based value if its consequences are aligned with that value. In a biomedical context, the degree of alignment with such values is measured by the amount of utility a given action provides to the patient. This category includes the values of *beneficence* (measuring positive utility, goods) and *non-maleficence* (measuring negative utility, harm).
- **Duty-based values:** An action is aligned with a duty-based value if and only if it is morally acceptable according to that value, regardless of its consequences. In a biomedical context, actions such as "cheating" or "killing a patient" would not be morally acceptable under any circumstance with respect to the duty-based value of *autonomy*.

Table 2: Doctors' actions.

Action	Description
Action 1: RCP	Cardiopulmonary Resuscitation (CPR)
Action 2: Transplant	Transplant
Action 3: UCI	Intensive Care Unit (ICU)
Action 4: VMNI	Non invasive respiratory support. Any form of ventilation support without tracheal intubation (includes oxigenotherapy)
Action 5: DVA	Vasoactive drugs (noradrenaline, epinefrine, dopamine...)
Action 6: Dialysis	Dialysis
Action 7: Simple: RX / Anal / Culture / AB	X-ray, blood analysis, cell or urine culture
Action 8: Mild: TC / Transf / Picc / Enteral nutrition	CAT Scan (computed tomography) / Blood transfusion / PICC: Peripheral Inserted Central Catheter / Enteral Nutrition
Action 9: ADV: RNM / Endoscopy / Parenteral nutrition	Nuclear Magnetic Resonance / Endoscopy / Parenteral nutrition
Action 10: Palliative surgery	Surgery designed to remedy the discomfort of and pain symptoms of incurable diseases. Palliative surgical procedures are intended to reduce suffering or support quality of life.
Action 11: Curative surgery	Are intended to prolong life or cure disease.

Table 3: Important values (some patient values are related to basic values).

Value	Value type	Description
Value 1: Autonomy	Basic value	patient's ability to make informed decisions over themselves
Value 2: Beneficence	Basic value	patient's benefit ensured
Value 3: Non-maleficence	Basic value	no harm being inflicted on the patient
Value 4: Justice	Basic value	fair, equitable, and appropriate treatment of all patients
Value 5: Cost efficiency	Hospital value	Cost minimisation
Value 6: Symptoms control (no pain)	Patient value	Minimising patients' suffering from pain
Value 7: Better quality of life	Patient value	Improving the quality life of patients

This category also includes the value of *justice*.

We can formalise alignment with a consequence-based value by considering a patient's medical conditions C before performing a medical action and comparing them with their medical conditions C' after the action is performed. Formally, let V be a consequence-based value, then:

$$align(a, \langle C, C' \rangle, V) = f_V(C, C'),$$

where a is the medical action taken, and f_V is a function comparing the two medical conditions.

There are two implications from this equation. The first one is that the action taken is irrelevant to the formula since we only care about the consequences. Moreover, this function is taking into account that the outcome of an action is non-deterministic in a medical context, and for that reason we must focus on its consequences.

The second implication is that we can obtain a formal definition of f_V (and thus, of the value) by explicitly listing which and how patient criteria in Table 1 are considered relevant. Assume that we have already agreed on the subset of criteria C_V associated with a given value V. Then, a possible formula for f_V could be:

$$f_V(C_V, C'_V) = G_V(\sum_{i=1}^{|C_V|} g_V^i(c_i, c'_i)),$$

where $c_i \in C_V$ and $c'_i \in C'_V$ are the conditions of the patient for each criterion i before and after the action, and functions g_v^i and G_V could be for instance $g_v^i(x,y) = x - y$ and $G_V(x) = x$. A formal definition of them would allow us to obtain a representation of beneficence and non-maleficence.

Finally, further research must be conducted for the other two values to assess which is the set of acceptable actions associated with each duty-based value.

4.2 Machine Learning

4.2.1 Value-Alignment of Actions in Context

One approach that we will investigate is to develop learning mechanisms that would use the corpus being built to help us learn and predict the relations between an action A, context C (defined through the patient criteria) and value V. In other words, we want to learn to answer the following question: *In a given context C, does an action A promote, demote, or not-affect a given value V?*

Formally, we specify these relations as the degree of alignment that the action A in context C has with the value V, which we represent as $align(A,C,V)$. We want the range of alignment to be $[-1,1]$, so that positive alignment would represent the action promoting the value, negative alignment would represent the action demoting the value, and an alignment of zero would represent the action not to affect the value. Furthermore, the use of a range helps us describe varying degrees of (mis)alignment.

We expect the model to predict the degree of (mis)alignment of an action with a value in a given context by learning from the past alignments presented in the corpus we are building. That is, the model learns from the past judgements of the medical personnel populating the corpus. Naturally, one of the main challenges of this approach will be the scarcity of the data.

4.2.2 Value-Alignment of Norms

We follow the traditional approach of defining norms through deontic operators over actions in context (Andrighetto et al., 2013; Ågotnes et al., 2009). Some examples of norms are:

- It is prohibited to perform action A (or action $\neg A$) in context C.
- It is permitted to perform action A (or action $\neg A$) in context C.
- It is obligatory to perform action A (or action $\neg A$) in context C.

These examples make use of three deontic operators: prohibitions, permissions, and obligations. We note that there are other deontic operators that one may consider, such as gratuitousness (permission to not perform an action) or indifference (permission to perform as well as to not perform an action), to name a few. We choose the above three operators for their common usage. Furthermore, it is well established that any deontic operator can be chosen as a basic operator, and then all other deontic operators can be defined in terms of the chosen basic operator. For example, a permission to perform an action can be specified as the negation of an obligation to not perform that action. As such, we say other deontic operators may easily be added, if the need arises, as they can be defined in terms of any of those three operators above.

Formally, we say let $N = D(A,C)$ specify a norm describing a deontic operator $D \in \{F,P,O\}$ (where F describes what is forbidden, P what is permitted, and O what is obligatory) over an action A when the context C is satisfied (or holds).

We then argue that if we can learn the alignment of an action A in context C with a value V, then we can deduce whether a norm $N = D(A,C)$ is aligned or not with that value V, which we represent as $alignN(N,V)$.

Examples of the properties that should hold when deducing the alignment of norms from the alignment of actions are presented in Figure 1. For example, we say that if an action A in a given context C is aligned with a value V (alignment is positive), and the norm states that this action A is permitted or obligatory in the context C, then this norm is aligned with that value because it permits (in the case of the deontic operator P) or obliges (in the case of the deontic operator O) the action A that promotes that value V. Furthermore, the alignment of the obligation may be greater than the alignment of the permission, since obligations are stronger than permissions in bringing about a given action. Similarly, if the norm prohibits the action A in context C (alignment is negative), then the norm is not aligned with the value V. Similar reasoning is followed in the remaining cases.

Finally, we note that we can also compute the alignment of one norm N with a set of values $\mathcal{V}$ by aggregating the alignment of that norm N with each of the values $V \in \mathcal{V}$:

$$alignN(N,\mathcal{V}) = \bigoplus_{V\in\mathcal{V}} alignN(N,V)$$

where $\bigoplus$ is an aggregation operator to be designed.

4.2.3 Value-Alignment of Protocols

Since we understand protocols as sets of norms, we say let $P = \mathcal{N}$ specify a protocol composed by the set of norms $\mathcal{N}$. We note that as our work progresses, we may need to modify our specification of protocols in such a way that allows us to address conflicting norms. For example, we may attach a priority measure to each norm, so that norms with higher priority can override norms with lower priority when conflicts arise. However, this requires further collaboration with the medical personnel at Hospital del Mar to confirm that whatever specification we use for protocols is consistent with their definition of protocols and how they deal with conflicting norms. For the time being, we keep things simple by defining protocols as sets of norms.

We then argue that if we can assess the alignment of norms with values, then we may deduce from that the alignment of protocols with values, which we represent as $alignP(P,V)$.

For example, by aggregating the alignment of each norm N in the set of norms $\mathcal{N}$ with value V, we get the alignment of the protocol $P = \mathcal{N}$:

$$alignP(P,V) = \bigoplus_{N\in\mathcal{N}} alignN(N,V)$$

$$
\begin{array}{ll}
IF & align(A,C,V) > 0 \\
THEN & alignN(O(A,C),V) \geq alignN(P(A,C),V) > 0 \wedge \\
 & alignN(F(A,C),V) < 0 \\
 & \\
IF & align(A,C,V) < 0 \\
THEN & alignN(O(A,C),V) \leq alignN(P(A,C),V) < 0 \wedge \\
 & alignN(F(A,C),V) > 0 \\
 & \\
IF & align(A,C,V) = 0 \\
THEN & alignN(O(A,C),V) = alignN(P(A,C),V) = alignN(F(A,C),V) = 0
\end{array}
$$

Figure 1: Properties of deducing the value-alignment of norms from the value-alignment of actions in context.

Similarly, we can also compute the alignment of a protocol P with a set of values $\mathcal{V}$ by aggregating the alignment of that protocol with each of the values $V \in \mathcal{V}$:

$$alignP(P,\mathcal{V}) = \bigoplus_{V \in \mathcal{V}} alignP(P,V)$$

Again, the design of the aggregation operator $\bigoplus$ will be carried out in future work.

4.3 Symbolic Reasoning

The alternative approach to learning that we will investigate is providing a symbolic representation of actions and values similar to (Montes and Sierra, 2022; Sierra et al., 2021; Osman and d'Inverno, 2023), and using that symbolic representation to reason about the alignment of actions in context with values. We describe next the requirements for the symbolic representations.

First, the symbolic representation of actions must define what are the outcomes of actions. That is, how do they change the current state of the world. In the NIT use case, actions change some of the patients' criteria, such as their expected survival, their comfort (for example, if an action lowers the fever of a patient), etc.

The symbolic representation of values, on the other hand, essentially defines how a value may be evaluated in a given context to assess whether it is being promoted or not. Section 4.1 has presented our initial approach.

For example, to evaluate whether the value 'non-maleficience' is promoted, which is understood as no harm is inflicted on the patient, the medical personnel try to confirm whether the patient has no pain, improved quality of life, and improved expected survival, to name a few. Those are all deduced from the criteria, such as the 'expected survival (in months)' and the 'Frail-VIG' value. As such, when populating the corpus and deciding whether a given action promotes or not a given value, the medical personnel are in fact analysing the impact of that action on those specific criteria, and deciding accordingly whether the value 'non-maleficience' is being promoted, demoted, or unaffected by that action.

Given the symbolic representations of actions and values, we can then develop a model that could analyse the changes that an action brings about (in our use case, that would be changes in patients criteria), and whether those changes result in promoting, demoting or not affecting a given value (as value evaluation is based on analysing patient criteria).

Similar to the machine learning approach, we argue that if we can reason about actions' alignment with values per context, then we can deduce norms alignment with values as well as the alignment of entire protocols with values, following the approaches presented in Sections 4.2.3 and 4.2.2.

4.4 Impact of Value Based Analysis

In both the machine learning approach and the symbolic reasoning approach, we can analyse the alignment of actions in context, norms, and even entire medical protocols with values. But what is the actual impact of this work? We argue that this can support value-aware decision-making for both the medical personnel and the management teams at hospitals as follows:

- By analysing the alignment of actions in context, the machine can inform a medical personnel whether the action they have decided to carry out is aligned or not with certain values, or whether this action prefers one value over another, and to what degree. We stress that we do not make statements about what is right and wrong. The model simply analyses the degree of (mis)alignment with values, and it is up to the medical personnel to evaluate such alignments as well as asses the importance of different values and make their decisions accordingly. In summary, this line of work promotes value-aware decision making by medical personnel.

- By analysing the alignment of norms and protocols, the machine can inform the management team (whether at the NIT level, or hospital level), when certain norms or protocols are aligned with certain values, or whether they give preference to one value over another, and to what degree. This helps the management team decide when, and under what condition, should norms and protocols change and evolve. Again, we do not make statements about what is right and wrong, but we promote value-aware protocol design and specification.

5 CONCLUSIONS

In this position paper, we have described the initial work on developing value aware AI and applying it to the medical field. We have described the process of selecting an appropriate and illustrative medical protocol to work with, the ongoing building of the relevant corpus, and the plans on how to develop models and mechanisms that would promote value-aware decision-making and value-aware protocol design and specification.

Our ongoing work continues to build the corpus in collaboration with Hospital del Mar. We will also continue with the formal specification of values, and commence the development of models and mechanisms that reason about the alignment of actions, norms, and protocols with values according to the plans presented in Section 4.

ACKNOWLEDGEMENTS

This work has been supported by the EU funded VALAWAI (# 101070930) project and the Spanish funded VAE (# TED2021-131295B-C31) and Rhymas (# PID2020-113594RB-100) projects. Furthermore, partial financial support was received from project 2021 SGR 00754.

REFERENCES

Ågotnes, T., van der Hoek, W., Rodríguez-Aguilar, J. A., Sierra, C., and Wooldridge, M. J. (2009). A temporal logic of normative systems. In Makinson, D., Malinowski, J., and Wansing, H., editors, *Towards Mathematical Philosophy*, volume 28 of *Trends in logic*, pages 69–106. Springer.

Andrighetto, G., Governatori, G., Noriega, P., and van der Torre, L. W. N., editors (2013). *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Askitopoulou, H. and Vgontzas, A. N. (2018). The relevance of the hippocratic oath to the ethical and moral values of contemporary medicine. part i: The hippocratic oath from antiquity to modern times. *European spine journal*, 27(7):1481–1490.

Beauchamp, T. and Childress, J. (1979). *Principles of Biomedical Ethics*. Oxford University Press.

Curto, G., Montes, N., Sierra, C., Osman, N., and Comim, F. (2022). A norm optimisation approach to sdgs: tackling poverty by acting on discrimination. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5228–5235. ijcai.org.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

McCarthy, J. (1999). Making robots conscious of their mental states. In *Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, July 1995]*, page 3–17, GBR. Oxford University.

Montes, N. and Sierra, C. (2021). Value-guided synthesis of parametric normative systems. In Dignum, F., Lomuscio, A., Endriss, U., and Nowé, A., editors, *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, pages 907–915. ACM.

Montes, N. and Sierra, C. (2022). Synthesis and properties of optimally value-aligned normative systems. *J. Artif. Intell. Res.*, 74:1739–1774.

Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. MIT press.

Ngan, O. M. Y. and Sim, J. H. (2021). Evolution of bioethics education in the medical programme: a tale of two medical schools. *International Journal of Ethics Education*, 6(1):37–50.

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

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.

Rodriguez-Soto, M., Serramia, M., López-Sánchez, M., Rodriguez-Aguilar, J. A., Bistaffa, F., Boddington, P., Wooldridge, M., and Ansotegui, C. (2023). Encoding ethics to compute value-aligned norms. *Minds and Machines*.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group.

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

Veatch, R. M. (2020). Reconciling Lists of Principles in Bioethics. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 45(4-5):540–559.