User Study Design for Identifying the Semantics of Bioethical Principles

Manel Rodriguez-Soto[1][0000−0003−1339−2018], Nardine Osman[1][0000−0002−2766−3475], Carles Sierra[1][0000−0003−0839−6233], Nieves Montes[2][0000−0001−6981−7893], Jordi Martinez Roldan[3], Rocio Cintas Garcia[4], Cristina Farriols Danes[4], Montserrat Garcia Retortillo[4], and Silvia Minguez Maso[4]

[1] Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain
{manel.rodriguez,nardine,sierra}@iiia.csic.es
[2] Mentice, Barcelona, Spain
[3] Hospital Sant Joan de Déu, Barcelona, Spain
[4] Hospital del Mar Research Institute (IMIM), Barcelona, Spain

Abstract. The topic of value alignment in AI has been gaining significant attention. The ultimate objective is how AI systems can align with human values. One of the main challenges, however, is identifying the relevant values. Some emerging works are focusing on learning relevant values through analysing our statements on social media. In most of these cases, learned values are simply specified through a label such as equality, fairness, or justice. However, the semantics of these values is not studied further. Addressing this gap, we propose a user study that provides us with a systematic approach for learning value semantics. The study is in the context of healthcare and the values of interest are the four fundamental bioethical principles: beneficence, non-maleficence, autonomy, and justice. We conduct a user study to collect the views of medical professionals on the value alignment of synthetic patient cases. We then search the space of candidate functions to find the one that best fits the participants answers, and hence, best describes their view of the value in question.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

The topic of value alignment is gaining a lot of traction lately [6, 11, 8]. New research lines are emerging that address issues such as how individual values can be aggregated to the level of groups [12]; how arguments that explicitly reference values can be made [3]; how decision making can be value-driven [21, 7, 9]; and how norms are selected to maximise value alignment [19, 14, 20]. One of the important challenges, however, is identifying the relevant values that AI needs to align with. There is a growing field on how AI learns human values, with many focusing on analysing our statements on social media [13, 4]. In almost all

these cases, learned values are simply specified through a label such as *equality*, *fairness*, or *justice*. However, the semantics of these values is not studied further. Recently, we have proposed a formal model for value representation that aims at tackling the issue of value semantics [15]. Values, in this proposal, are no longer treated as labels representing abstract concepts, but as complex taxonomies that take into consideration the relations between value concepts, the importance of value concepts, and the semantics of value concepts. Semantics are provided by linking abstract concepts with formal properties whose satisfaction can be measured. Value semantics essentially define what it means for a given behaviour, whether a single action or an entire interaction, to be aligned with a given value.

However, the issue of learning value semantics arises. Relying on humans to provide these semantics is not always reliable. This paper explains our experience in trying to define the four basic bioethical principle —beneficiency, non-malificiency, autonomy, and justice— and our current user study that is being unrolled to achieve a more systematic approach for extracting these semantics.

Our initial approach was fruitful, but time consuming and not very efficient in eliciting the views of a large number of medical professionals. In our initial approach, our attempt to formalise the semantics of these four bioethical principles was realised through engaging in intensive and lengthy discussions with three medical doctors from Hospital del Mar, Barcelona (the last three authors of this paper). Our interactions, which spanned a period of several months in 2023–2024, allowed us to come up with clear equations for two of the four bioethical principles: beneficency and non-maleficency. However, this labour intensive work allowed us to provide a formal specification that reflects the views of three medical doctors only.

Our current, more systematic, approach is to use replicable questionnaires that can be easily conducted in hospitals at a large scale, allowing us to extract the semantics of these four bioethical principles by taking the point of view of a larger number of medical professionals, both doctors and nurses. By achieving an easily replicable study, this opens the door to comparing the semantics of these bioethical principle for different hospitals, different countries, or even different cultures.

The rest of this paper is divided as follows. Section 2 introduces the four bioethical principles according to the literature of biomedical ethics. Section 3 presents our initial work in defining the semantics of these principles. Section 4 presents our ongoing user study for extracting the semantics of these principles at a large scale, followed by some concluding remarks in Section 5.

2 The Four Bioethical Principles

This paper considers that there are four main bioethical principles, following Beauchamp and Childress' principialism [1]. As agreed upon by the biomedical community, at least these four principles provide the best framework for ethical analysis in biomedical scenarios [22]. These four principles are:

1. **Autonomy:** to give value to the considerations and options of autonomous patients, and refrain from putting obstacles to their actions, unless they are clearly harmful to others.
2. **Beneficence:** to ensure that the benefit of the patient is always maximised.
3. **Non-maleficence:** to ensure that no harm is being inflicted on the patient or at least that it is being minimized.
4. **Justice:** to proceed with equity in the distribution of burdens and benefits and treat everyone fairly, equally, and without discrimination.

These principles are sometimes referred to as medical ethical values in the literature [16, 17], and as such, we use the words principles and values interchangeably in this paper.

Following Ross's ethical theory, all four values are *prima facie* [18]. That is, there is no explicit ordering between them, and ideally a clinicial should always try to behave as value-aligned as possible with respect to all four values. However, Ross (and later Beauchamp and Childress) acknowledged that always completely aligning with multiple conflicting values would be impossible. Thus, as *prima facie* values, one should always weigh the possible benefits and costs of each medical value according to each value and then decide which is the most appropriate one.

As an example, consider a patient that is suffering from a serious illness and requires an immediate surgery. Following the value of beneficence, the clinician should seek to maximise the benefit of the patient, which means to perform surgery on the patient. However, following the value of non-maleficence the clinician should never produce harm on the patient, or at least minimise the risk of harm. But any surgery has an inevitable degree of risk, and always causes a short-term harm to the patient. Under this dilemma, principialism recommends evaluating the advantages and disadvantages of either performing the surgery or not, in terms of the values in conflict. If the degree of risk of the surgery is low enough while the benefits are almost guaranteed, a clinician then should select to prioritise beneficence over non-maleficence in this specific case.

Next Section 3 present our first attempt at formalising these four biomedical values so that they are computationally treatable.

3 Initial Approach

Our first step towards formalising the four biomedical values was to categorise them following the proposed outline by Veatch in [22]. Veatch states that biomedical ethics' four main values can be divided into two categories: *consequence-based* values and *duty-based* values. Recall that value semantics essentially specify what it means for a given behaviour to be aligned with a given value. However, 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, good) 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 "misinforming the patient" or "disrespecting patient's directives" would not be morally acceptable under any circumstance with respect to the duty-based value of *autonomy*. This category also includes the value of *justice*.

As such, the approach for formalising each of the above value categories is different, as we present next.

3.1 Formalising Consequence-Based Values

We can formalise alignment with a consequence-based value by considering a patient's medical state, described by a set of criteria C, before performing a medical action and comparing it with their medical state after the action is performed, described by the change in criteria C'. Formally, let V be a consequence-based value, then:

$$align(a, \langle C, C' \rangle, V) = f_V(C, C') \tag{1}$$

where a is the medical action taken, and f_V is a function comparing the patient's medical state before and after the action a.

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 it evaluates) by explicitly listing which patient criteria are considered relevant for that specific value and how its change in criteria is evaluated. Assume that we have already agreed on the subset of criteria C_V relevant for a given value V. Then, a possible formula for f_V could be:

$$f_V(C_V, C'_V) = G_V(\sum_{i=1}^{|C_V|} g_V^i(c_i, c'_i)) \tag{2}$$

where $c_i \in C_V$ and $c'_i \in C'_V$ are different criteria describing the state of the patient before and after the action. Different definitions of the functions g_v^i and G_V could be explored. For example, we can have $g_v^i(x, y) = x - y$ and $G_V(x) = x$, which essentially states that we simply account for the accumulated change for each relevant criterion. If the change in a criterion's evolution was positive (numbers went down after the action was performed), then that would result in

a positive alignment, and hence, the change in that criterion's evolution would be contributing to promoting the value V, and vice versa. As such, we see that the range of each criterion needs to be designed and normalised accordingly.

Equation 2 provides us with the formal representation of the consequence-based values, beneficence and non-maleficence. What is then necessary to identify is the set of relevant criteria C_V for each of those values, and the exact definitions of g_v^i and G_V.

Our interactions with the three medical doctors from Hospital del Mar allowed us to identify the potentially relevant patient criteria for the values beneficence and non-maleficence. These are presented in Table 1.

Defining functions g_V and G_V for each value was a more complicated task, and required continued intensive discussions, which was very time consuming and inefficient, especially that the resulting functions would reflect the point of view of the few medical doctors one is working with. For this reason, a more systematic approach has been designed through user studies, which we present in the next section. The user study allows us to define the functions g_V and G_V in an efficient approach that takes into consideration the point of view of a large set of medical professionals, both doctors and nurses.

3.2 Formalising Duty-Based Values

Formalising the alignment of behaviour with a duty-based value is a much more complex task, it requires an understanding of the moral norms that define what is morally acceptable behaviour for that value. The alignment is then dependant on the action and the moral norms that define a duty-based value. Formally, let $\mathbf{V}$ be a duty-based value and $N_{\mathbf{V}}$ be the set of moral norms defining that value $\mathbf{V}$, then:

$$align(a, N_{\mathbf{V}}, \mathbf{V}) = f_{\mathbf{V}}(a, N_{\mathbf{V}}) \tag{3}$$

To compute this alignment, we say it must be based on the degree of satisfying those identified moral norms. So we have:

$$f_{\mathbf{V}}(a, N_{\mathbf{V}}) = \bigoplus_{n \in N_{\mathbf{V}}} (sat(a, n)) \tag{4}$$

where $sat(n)$ describes the degree of satisfaction of action a with moral norm n, and $\bigoplus$ describes some aggregation function that aggregates the different degrees of satisfaction for all moral norms of $\mathbf{V}$.

As such, it becomes critical to understand and formalise these moral norms that define a given value, along with defining the $\bigoplus$ and sat functions.

Our interactions with the medical doctors from Hospital del Mar allowed us to identify the moral norms associated with the value autonomy, which we present below:

norm 1 The patient must be informed of the implications of both receiving and not receiving a given treatment, unless they choose to remain uninformed.

norm 2 The patient's wishes concerning treatments must be respected.

norm 3 The patient must not be coerced or persuaded to accept or reject a treatment.

Our findings are relatively consistent with the literature on that topic [5]. In [5], two norms are represented by our first norm (norm 1): one declares the patient's right to be informed of the consequences of treatments, and the second declares the patient's right to choose not to be informed of treatments' consequences (or the right to remain uninformed). Three norms are represented by our second norm (norm 2): the first declares the right not to be subjected to any treatment without the consent of the patient, the second declares the patient's right to withdraw their consent at any time, and the third states that when the patient is not in a position to express their wishes then their latest wishes (advanced directives) must be respected. We believe our second norm summarises all three situations. Our third norm (norm 3) adds the requirement that a patient must not be coerced or persuaded to accept or reject a treatment, which also appears in the discourse of existing literature.

With the given moral norms identified, the next step is to define the $\bigoplus$ and *sat* functions. Concerning the *sat* function, we assume the results of the *sat* function to be known and provided by the patient's medical team. This is because the *sat* function, or the degree of satisfaction of a given action with given moral norm, depends on complex issues, such as understanding the informed consent forms (norm 1), understanding the wishes of the patient with respect to different treatments (norm 2), and identifying whether the patient has been coerced or not in accepting/rejecting a treatment (norm 3). All of this is not straightforward and requires the analysis of natural text in documents (such as the informed consent forms) and live interactions (such as the discussions that a doctor has with a patient and/or his family members). For this reason, in this paper we assume the degree of satisfaction (*sat*) of these norms to be provided. We rely on the medical professionals' capability of identifying and declaring the satisfaction of these norms. Table 2 presents the variables n_1–n_3, corresponding to the satisfaction of norms 1–3. For the time being, the range of satisfaction of a moral norm is either 1, to represent that the norm was respected, or 0, to represent that the norm was not respected. Future work can introduce degrees of satisfaction, as needed.

As for $\bigoplus$, one suggested implementation is $\bigoplus = \sum_{n \in N_{\mathbf{V}}} w_n \cdot sat(a, n)$, where the aggregation is a linear equation that assumes a linear relationship between the degree of satisfaction of the various moral norms, each with a predefined weight w_n.

Once again, trying to elicit the exact definition of $\bigoplus$ that captures the semantics of autonomy through focus groups is a complicated and time consuming task. We have provided here one possible proposal for $\bigoplus$. We hope that the user study we are designing will allow us to capture a more comprehensive definition of $\bigoplus$ that captures the point of view of a large set of medical professionals from Hospital del Mar.

The value justice is much more complex and requires further work. The literature provides open ended discussions and definitions on the topic [5]. For example, it may be understood as ensuring fair treatment and access to resources for all, regardless of their social or economic status. In general, it is about avoiding unjust discrimination based on factors such as race, ethnicity, gender, religion, socioeconomic status, disability, or age. But it may also touch upon topics like holding individuals and institutions responsible for their actions and decisions. What is clear is that justice is the only value that is not solely concerned with the patient in question, but third parties as well. At this stage, no clear moral norms have been identified for the value justice. Through the setup of our user study (the corpus building stage of Section 4.2), we will aim at identifying and eliciting moral norms that are relevant for justice. If we do succeed, then a follow up user study can be designed to investigate the functions $\bigoplus$ and *sat* defining justice for the identified set of moral norms.

4 Proposed User Study

This section proposes a user study for identifying the semantics of the four biomedical values, helping us provide a concrete computational definition of these values that captures the point of view of a large group of medical professionals. In other words, we essentially aim to identify Equations 1 and 3 that provide a best representation of the point of view of the medical professionals involved in our study. First, Section 4.1 discusses the issue of identifying relevant variables. This step is crucial as these are the variables upon which the equations giving values their semantics will be built upon. Then, Section 4.2 illustrates the various stages of our proposed user study and how its results will be used to craft the equations defining the four bioethical values, building on the identified variables.

4.1 Variable Identification

One main requirement for formally defining the formal expression that gives any value its semantics is recognising the potential variables that this expression is built upon. As an example, for formally defining income inequality, the Gini index [10] depends on two variables: the income or wealth values of individuals/households, and the population size (total number of individuals/households). As such, to formally define our four biomedical values, the first step is to identify the relevant variables that define each value. Our initial approach (Section 3) has paved the ground for identifying these variables. Interactions with the stakeholders are necessary at this stage.

Given the definitions of beneficence and non-maleficence —ensuring patient benefit is maximised and ensuring no harm is inflicted on the patient [1]—, it becomes clear that the relevant variables for those two bioethical values should be those describing the patient's state, as benefit and harm can be assessed by evaluating the change in the patient's state. However, we cannot depend on

Variable Name	Variable Description	Variable Range
c_1	**Age:** Specifies the patient's age, specified in intervals representing decades.	{0–19, 20–29, . . . ,90–99, +99}
c_2	**Complex Chronic Disease (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.	{Yes, No}
c_3	**Short-term survival (MACA):** Measures if the patient has an advanced chronic disease with an expected survival rate of less than 12–18 months that requires palliative care.	{Yes, No}
c_4	**Expected survival:** Provides an estimation, in months, of the expected survival of the patient.	{$<$ 12 months, $>$ 12 months}
c_5	**Frailty Index (Frail-VIG):** Assesses the degree of frailty of the patient.	{Low, Moderate, High}
c_6	**Clinical Risk Group (CRG):** Provides a categorical classification that uses administrative data to identify patients with chronic health conditions.	{0, 1, 2, 3, 4}
c_7	**Social Support:** Specifies if the patient has social support (from family or friends) to offer support functions (emotional, instrumental, ...)	{Yes, No}
c_8	**Functional independence (Barthel Index):** Measures the capacity of the patient with respect to executing activities of daily living (ADL), such as feeding, bathing, ambulation, bladder and bowel control, ...	{0–20%, 21–60%, 61–90%, 91–99%, 100%}
c_8	**Instrumental activities independence (Lawton Index):** Measures an individual's ability to perform various complex activities that are necessary for independent living, such as using the telephone, shopping, food preparation, housekeeping, laundry, mode of transportation, responsibility for own medications, managing finances, ...	{0, . . . , 8}
c_9	**Patient's advanced directives:** Specifies, for patients with a decision-making capacity, if there is a signed document or oral communication describing the patient's desires regarding treatment decisions. This includes when the patients identify whom they want to make decisions on their behalf when they cannot do so themselves.	{Yes, No}
c_{10}	**Cognitive deterioration:** Specifies if the patient suffers cognitive impairment, such as confusion, memory loss, difficulty understanding or speaking, problems with concentration, ...	{No deterioration, Low-Moderate, Severe }
c_{11}	**Emotional state:** Specifies whether the patient suffers emotional distress.	{Yes, No}
c_{12}	**Discomfort level:** Measures level of pain and physical distress.	{Low, Medium, High}

Table 1. The identified generic (patient) variables related to beneficence and non-maleficence

variables that are too specific for a given context, such as a given diagnosis or a given field in medicine. For this reason, we set out, in collaboration with the three medical doctors from Hospital del Mar, to identify generic patient variables that can be applied to patients regardless of context, and yet are informative enough to allow medical professionals perform a practical and effective analysis of the patient's state and its evolution. These generic patient variables are listed in

Variable Name	Variable Description	Variable Range
n_1	The patient's (or their authorised representative's) level of understanding of the instructions provided by the medical professionals.	{0,1}
n_2	The patient (or their authorised representative) being informed about each possible treatment and the consequences of receiving or not that treatment.	{0,1}
n_3	The patient's (or their authorised representative's) independent decision (that s/he has not been coerced/pressured by a third party to accept or reject a given treatment).	{0,1}

Table 2. The identified generic variables related to autonomy

Table 1. These variables play a vital role in the definition of consequence-based values, as Equation 1 shows. The new equations that we intend to identify that capture the point of view of a large group of medical professionals will also build upon these criteria, hence the need for interacting with stakeholders ahead of the user study to identify such criteria. We also hope that qualitative feedback from the user study can confirm whether our identified criteria are comprehensive or lacking in some respects.

Regarding the other two duty-based values (autonomy and justice), and as we noted in Section 3, these depend on moral norms. As such, it is necessary to identify and formalise the moral norms that define them.

Given the definition of autonomy, based on some literature on the topic [1, 2, 5] and our discussions with the medical doctors from Hospital del Mar, we were able to recognise the moral norms that define autonomy and the three variables that describe the degree of satisfaction of those norms are presented in Table 2. The equation defining the value of autonomy should be built on these three variables. But we also look forward to qualitative feedback to confirm whether our selected moral norms are comprehensive or lacking in some respects.

Given the definition of justice, we note that justice is even more complicated than autonomy, as alignment with justice cannot be computed at the level of the individual. For example, to assess whether a given action towards a patient is aligned with the value of justice, one of the requirements is to compare it with the actions that similar patients have received. Hence, due to the broadness of the definition of justice, we were not able to identify variables of justice to be included in our user study, but we hope the user study would shed light on the moral norms that define justice.

4.2 Function Identification

The next step, after identifying the potential relevant variables through interacting with the stakeholders, is to identify the functions that define the bioethical values through these variables: function f_v of Equations 1 and 3.

Instead of directly asking stakeholders for their views on these bioethical values and how can we define such functions, as in Section 3, we design a user study that would elicit these views through simple questionnaires. This allows us to easily replicate this user study in different departments, hospitals, and so on, to get the different perspectives on the semantics of these bioethical values. In essence, the questionnaires present medical professionals with different patient cases and asks whether the medical professional believes each of the bioethical values was promoted, demoted or not affected for each of those patient cases. The idea is that it is easier for medical professionals to give their opinion on the promotion/demotion of a value when presented with a concrete case. And with these responses, we can then search for the function that best represents the medical professionals' answers; that is, find the function that best describes their perspectives.

The steps for conducting this user study are as follows:

1. **Build a diverse corpus of patient cases**, to be used in the user study.
2. **Select the participants (medical professionals)** for conducting the user study.
3. **Conduct the user study** by asking the selected medical professionals to assess whether each of the bioethical pricinples was promoted, demoted, or not affected for each of the patients cases they are presented with.
4. **Identify the functions** that describe the semantics of the various bioethical values. This is achieved by finding the functions that best fits the responses of medical professionals.

The remainder of this section is structured with each subsequent subsection covering the different steps of our user study.

Building the Corpus Obtaining real data is usually a challenge in the medical field, especially when the identified relevant variables refer to information that is usually not digitised in many hospitals. As such, for our user study, we chose to build a corpus of synthetic data.

The objective is to compile a diverse and large number of patient cases, where the patient is defined by his/her relevant variables (in our case, generic patient variables and autonomy-related variables), along with the patient's diagnosis. The patient's diagnosis is concise and described in a few words only, and it is added simply to provide the context for medical professionals to help them with the assessment of patients in the first phase of the user study, when predicting the evolution of a patient's state (see Section 4.2). The diagnosis is not relevant (and not presented) in the second phase of the user study, when the promotion/demotion of values is assessed.

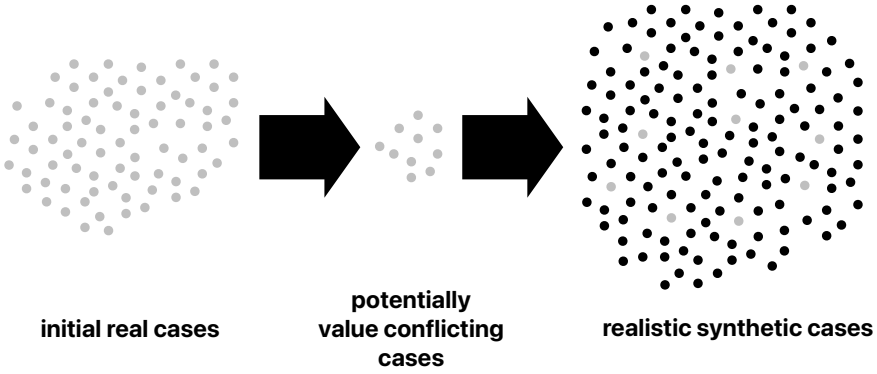

Fig. 1. Synthesising patient data

To minimise the number of patient cases that each medical professional has to assess in the user study (Step 4), we focus on cases that have the potential of raising a value conflict. In other words, cases that promote one bioethical value while demoting another. We believe such cases might be the most informative ones when it comes to learning the semantics of bioethical values from them.

To create realistic synthetic data, we choose to start with 75 actual patient cases. We are asking five of the medical professionals that we have been collaborating with to assess 15 cases each. They are asked to: 1) identify the cases that might potentially raise value conflicts, 2) identify which actions are the ones that might result in value conflicts, and 3) for each of those cases and their corresponding actions, identify the relevant patient criteria that they believe impact having a value promoted/demoted. In summary, this analysis helps us identify the cases with potential value conflicts, and the relevant actions and patient criteria associated with that potential value conflict. Furthermore, allowing for open-ended responses at this stage allows us to better understand if our list of criteria is comprehensive or not.

With these cases at hand, we plan to generate a larger number of realistic synthetic cases (we are aiming at 150 cases) that will eventually be used in the user study (Step 3). To generate these cases, we essentially take each of the potentially conflicting cases identified above and vary some of their identified relevant variables to create new cases. These new cases will allows us, through the user study, to pinpoint how the relevant variables may impact value promotion/demotion, from which we can then try to identify the functions defining the semantics of the four bioethical values (Step 4).

Figure 1 provides us with an illustration on how the 150 patient cases used in the user study are created.

An important issue to note is the need to create cases that span diverse diagnoses. This could help verify whether value promotion/demotion is only affected by the identified relevant variables, or whether the diagnosis plays a crucial role.

Our hypothesis, along with that of the medical doctors we are collaborating with, is that the diagnoses is not relevant when considering bioethical values: only the identified relevant variables impact value promotion/demotion. This hypothesis arises despite the fact that doctors are trained to always look at the diagnosis whenever assessing a given patient.

Selecting the Participants The choice of medical professionals to participate in the user study depends on whose perspective are we interested in obtaining. The semantics of bioethical values can be obtained, for example, for:

- Individual medical professionals, by asking each medical professional to assess a large number of diverse patient examples, which will allow us to extract the function that best defines their view concerning the four bioethical values.
- A hospital department, by asking a number of medical professionals of that department to assess a large number of diverse patient examples.
- A hospital, by having medical professionals from across the various departments of this hospital assess a large number of diverse patient examples.

Of course, we can proceed with this approach to assess the semantics for a given country or geographical region. We can also include and compare the perspective of nurses versus medical doctors.

Our current plans are to include nurses and medical doctors from a range of departments at Hospital del Mar, Barcelona. We plan to recruit 10 medical professionals in totals, with each being asked to assess 15 patient cases from the 150 cases generated in Step 2.

Conducting the User Study The user study is conducted in two phases. In the first one, participants (selected medical professionals) are presented with a number of patient cases: 15 per professional. Each case consists of patient criteria (variables of Table 1) and a selected action (identified in the earlier stage as potentially raising value conflicts). In this phase, each patient case is being assessed by one professional. The participants are asked to predict the change in patient criteria (specifically those of Table 1) when considering the selected action. At this stage, the diagnosis is presented, as the diagnosis is critical for predicting the evolution of a patient's state. Figure 2 provides an illustration of the questionnaire used in this phase. The parts in red highlight what needs to be filled in by the participants. In this case, how patient criteria evolve. As such, the results of these questionnaires provide us with an expectation of how some of the patients' criteria would evolve for each patient case and selected action.

After the first phase is completed, the second phase starts. In this phase, each participant is presented with 45 patient cases, where each case includes the patient criteria (variables of Tables 1 and 2), a selected action (with potential of introducing value conflicts), and some of the patient criteria (variables of Table 1) resulting from performing the presented action. The diagnosis is not

presented in this phase. Here, each patient case is being assessed by three medical professionals. We make sure here that the patient cases presented at this stage to a given participant are different from those assessed by that participant in the first phase. This ensures that knowing the diagnosis and thinking about the evolution of criteria does not influence one's views with respect to value alignment. In this phase, participants are asked to assess whether each of the bioethical values was promoted, demoted, or not affected for each of the patient cases they are presented with. Figure 3 provides an illustration of the questionnaire used at this stage. Again, the parts in red highlight what needs to be filled in by the participants. In this case, which values are being promoted/demoted. The results of these questionnaires essentially contain information on how the generic variables (both patient and autonomy related variables identified in Tables 1 and 2) impact value promotion/demotion. The next step will be to find the formal function for each value that fits the data, that is, the function that results in value promotion/demotion that fits the data.

Identifying the Semantics of Bioethical Principles This subsection describes how we plan to identify and formally define the semantics of a given bioethical principle, or value. The objective is to find the function f_v for a given value v, whether consequence-based or duty-based, that reflects the answers of the participants for that value v.

The search for function f_v can be framed as an optimisation problem over the space of functions $\mathcal{F} : \mathbb{Z}^2 \rightarrow \mathbb{R}$ that map a tuple of variables $X = (x_1, ..., x_n, r_v)$ to a real number.

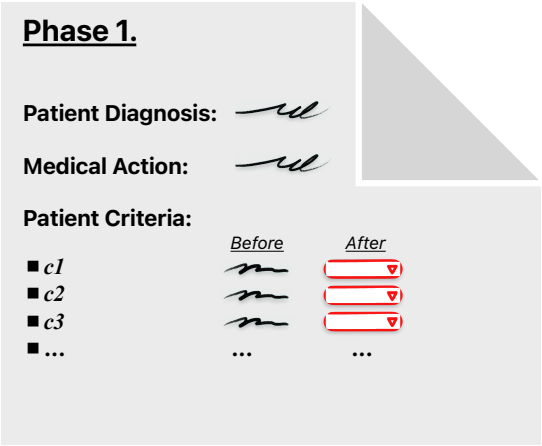

Fig. 2. The user study's questionnaire, phase 1

The variables (excluding the last variable r_v) represent the identified relevant criteria for a given value (Tables 1 and 2). In other words, if we are looking for the functions f_b and f_{nm} that define beneficence and non-maleficence, respectively, then we will have $X = (c_1, \ldots, c_n, c'_1, \ldots, c'_n, r_v)$; whereas, if we are looking for a function f_a that defines the value autonomy, then we will have $X = (n_1, n_2, n_3, r_v)$. Recall that r_v is the participants assessment, for a given patient case, on whether the value in question is promoted, demoted, or unaffected.

The objective is to find the expression $f_v(X)$ that best predicts the observed answers that participants have taken. To solve the problem, we implement an Evolutionary Strategy search algorithm. To generate potential candidates for $f(\cdot)$, we start with a simple grammar of arithmetic expressions:

$$\begin{aligned} Arg &:= x_1 \mid \ldots \mid x_n \mid r_v \mid z \\ Op &:= + \mid - \mid \cdot \mid / \\ Exp &:= Arg[OpExp] \end{aligned}$$

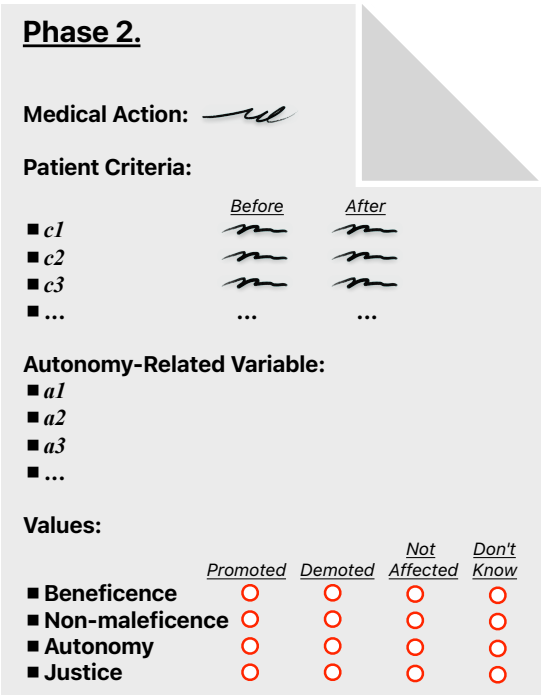

Fig. 3. The user study's questionnaire, phase 2

where $\{x_1, \ldots, x_n\}$ is the set of variables identified as relevant to the value in question, $r_v = \{-1, 0, 1\}$ stands for the participant's opinion on whether the value v is promoted ($r_v = 1$), demoted ($r_v = -1$), or unaffected ($r_v = 0$), and $z \in \mathbb{Z}$ is an integer that may be used when constructing possible candidate functions. For the time being, to keep the search space limited, we stick to a simple language that only makes use of the basic arithmetic operators: addition ($+$), subtraction ($-$), multiplication ($\cdot$) and division ($/$). Future work can incorporate additional operators, such as the *max* and *min* operators.

To evaluate how well does a candidate function $f(\cdot)$ fit the user study data, we consider two opposite factors. First, we assess what is the affinity of that function for the experimental data, which we define as $\mathit{Aff}(f)$. One approach for implementing $\mathit{Aff}(f)$ is through the mean absolute error, which computes the distance between the predicted value alignment f and the actual alignment presented in the responses of the user study data. Second, we assess how complex is that function. We favour simpler, less complex expressions. We define the complexity of an expression $\mathit{Comp}(f)$ as the number of operations (Op) required to generate it.

Our proposal then evaluates the suitability of a function by considering its affinity for the user study data and add a penalising factor for the complexity of the expression. The best function that describes the value in question is then obtained through optimisation:

$$f^* = \arg\max_{f \in F}(\mathit{Aff}(f) - (\lambda \cdot \mathit{Comp}(f))) \tag{5}$$

where λ controls the weight of penalisation and F is the set of functions that can be generated with the grammar.

The obtained function f^* that considers the participants' responses r_v for value v essentially describes the semantics of the value v. This could be computed taking into consideration the responses of one participant only, in which case the identified function f_v^* would represent the point of view of that participant. It could also consider the responses r_v of all participants, in which case the identified function f_v^* would represent the point of view of the collective of participants. This collective could be a department, a hospital, or even a geographical region. In our case, we find the function that best describes the point of view of the 10 medical professionals (doctors and nurses) recruited from various departments at Hospital del Mar.

5 Conclusion

With the rising interest in the role of human values in AI and the recent trend of learning these values from our online interactions [13, 4], we notice a striking lack of research on learning the semantics of these values. This paper aims to address this gap by proposing an approach that learns value semantics through user studies eliciting human feedback on value alignment. Our focus is on the medical field, and our participants are medical professionals who are presented

with various patient cases. For each case, they assess whether a given action promotes, demotes, or does not affect the value in question. We conducted a user study to evaluate the four bioethical values: beneficence, non-maleficence, autonomy, and justice. Based on the participants' responses, we then explore the space of candidate functions to find the one that best fits the user study answers regarding value alignment. In other words, we identify the function that best describes, computationally, what it means for an action (behaviour) to be aligned with each value.

One primary direction for future work is optimising the creation of synthetic data to improve our user study. Our current approach for building the corpus has been based on medical professional's collaboration for curating a number of cases (75 cases), and then eliciting the professionals' feedback for detecting potentially value conflicting cases. To help improve replicability, we are exploring the creation of synthetic data using ChatGPT. Ongoing work is promising, but naturally, expertise's opinion is still needed to help guide the generation of these cases. Though the involvement of the experts will be less time consuming.

If replicability is achieved, then we plan to apply this user study to a larger number of medical professionals, across different hospitals and geographical areas. This would allow us to analyse how value semantics change across departments, hospitals, or even countries.

Acknowledgments 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.

References

1. Beauchamp, T., Childress, J.: Principles of Biomedical Ethics. Oxford University Press (1979)
2. Beauchamp, T.: The 'Four Principles' Approach to Health Care Ethics, pp. 3 – 10. Wiley (06 2007). https://doi.org/10.1002/9780470510544.ch1
3. Bench-Capon, T., Atkinson, K.: Abstract argumentation and values. In: Simari, G., Rahwan, I. (eds.) Argumentation in Artificial Intelligence, pp. 45–64. Springer US, Boston, MA (2009). https://doi.org/10.1007/978-0-387-98197-0_3
4. Brugnoli, E., Gravino, P., Prevedello, G.: Moral values in social media for disinformation and hate speech analysis. In: Preproceedings of the Value Engineering in AI Workshop, at 26th European Conference on Artificial Intelligence (ECAI 2023) (2023)
5. Casado, M., Baroni, M.: Manual de bioética laica: Cuestiones de salud y biotecnología -. Colección de bioética, Universitat de Barcelona Edicions (2018), https://books.google.es/books?id=LfnBzgEACAAJ
6. 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)
7. Chhogyal, K., Nayak, A.C., Ghose, A., Dam, H.K.: A value-based trust assessment model for multi-agent systems. In: IJCAI. pp. 194–200. ijcai.org (2019)

8. European Comission: Ethics guidelines for trustworthy ai. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2019), accessed: 2021-06-29
9. Cranefield, S., Winikoff, M., Dignum, V., Dignum, F.: No pizza for you: Value-based plan selection in bdi agents. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 178–184. ijcai.org (2017). https://doi.org/10.24963/ijcai.2017/26, https://doi.org/10.24963/ijcai.2017/26
10. Damgaard, C.: Gini coefficient. https://mathworld.wolfram.com/GiniCoefficient.html, accessed:2024-06-11
11. IEEE: Ieee global initiative on ethics of autonomous and intelligent systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (2019), accessed: 2021-06-29
12. Lera-Leri, R., Bistaffa, F., Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A.: Towards pluralistic value alignment: Aggregating value systems through l_p-regression. In: Faliszewski, P., Mascardi, V., Pelachaud, C., Taylor, M.E. (eds.) 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022. pp. 780–788. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS) (2022). https://doi.org/10.5555/3535850.3535938, https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p780.pdf
13. Liscio, E., van der Meer, M., Siebert, L.C., Jonker, C.M., Mouter, N., Murukannaiah, P.K.: Axies: Identifying and evaluating context-specific values. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. p. 799–808. AAMAS '21, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2021)
14. Montes, N., Sierra, C.: Value-guided synthesis of parametric normative systems. In: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (eds.) AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021. pp. 907–915. ACM (2021)
15. Osman, N., d'Inverno, M.: A computational framework of human values. In: Dastani, M., Sichman, J.S., Alechina, N., Dignum, V. (eds.) Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024. pp. 1531–1539. ACM (2024). https://doi.org/10.5555/3635637.3663013, https://dl.acm.org/doi/10.5555/3635637.3663013
16. Page, K.: The four principles: Can they be measured and do they predict ethical decision making? BMC Medical Ethics **13**(1), 10 (May 2012). https://doi.org/10.1186/1472-6939-13-10, https://doi.org/10.1186/1472-6939-13-10
17. Rezler, A.G., Lambert, P., Obenshain, S.S., Schwartz, R.L., Gibson, J.M., Bennahum, D.A.: Professional decisions and ethical values in medical and law students. Academic Medicine **65**(9) (1990)
18. Ross, W.D.: The right and the good. Philosophy **6**(22), 236–240 (1930)
19. Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A.: A qualitative approach to composing value-aligned norm systems. In: Seghrouchni, A.E.F., Sukthankar, G., An, B., Yorke-Smith, N. (eds.) Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020. pp. 1233–1241. International Foundation for Autonomous Agents and Multiagent Systems (2020)

20. Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., Perelló, A.: Value alignment: a formal approach. CoRR **abs/2110.09240** (2021), https://arxiv.org/abs/2110.09240
21. di Tosto, G., Dignum, F.: Simulating social behaviour implementing agents endowed with values and drives. In: Giardini, F., Amblard, F. (eds.) Multi-Agent-Based Simulation XIII - International Workshop, MABS 2012, Valencia, Spain, June 4-8, 2012, Revised Selected Papers. Lecture Notes in Computer Science, vol. 7838, pp. 1–12. Springer (2012). https://doi.org/10.1007/978-3-642-38859-0_1, https://doi.org/10.1007/978-3-642-38859-0_1
22. Veatch, R.M.: 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 (07 2020). https://doi.org/10.1093/jmp/jhaa017, https://doi.org/10.1093/jmp/jhaa017