**Building an Algorithmic Advisory System for Moral Decision Making in the Health-Care Sector**

*Lukas J. Meier, University of Cambridge (United Kingdom)*

*Alice Hein,Technical University of Munich (Germany)*

*International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL*

**Keywords:** Artificial intelligence, medical ethics, computing ethics

**Extended Abstract**

Can medical ethics be made computable? Machine intelligence already helps health-care staff with a number of tasks. Ethical decision-making, however, has not been handed over to machines. We developed the world’s first algorithm for advising on a range of moral dilemma situations that occur in medical institutions.

In this talk, we explain how we chose an appropriate ethical framework, why we used fuzzy cognitive maps to set up the advisory system, and how we utilised machine learning to train it. We report on the difficult task of operationalising the principles of beneficence, nonmaleficence and patient autonomy, and describe how we selected suitable input parameters that we extracted from a training dataset of clinical cases.

**Ethical Basis**

Like any ethical judgments taken by humans, ethical algorithmic decision-making must be rooted in a moral framework. When constructing an artificial moral agent, the primary question is, therefore, that of the underlying normative ethical theory. Roughly speaking, philosophers distinguish between three fundamental types of ethical theories: the teleological, the deontological, and the aretaic.

According to teleological approaches, the *consequences* of an act determine whether the latter is morally right. Conversely, deontologists hold that actions are morally right when they conform to a particular *norm* or set of norms. The third fundamental normative ethical theory is virtue ethics. According to this framework, moral actions are the result of an individual’s acquiring praiseworthy dispositions of character.

Given the three theories’ complementary strengths and weaknesses, even hundreds of years of philosophising have not resulted in one of them emerging as superior. Until recently, this problem pertained only to human action. However, our ability to construct autonomous agents now requires that we make a decision regarding which moral principle to implement.

In the domain of clinical ethics, however, it became clear early on that consultants could not afford on a daily basis to engage in lengthy debates about which fundamental moral theory ought to prevail. Less general approaches of greater practical applicability were developed – among them casuistry, narrative ethics, feminist ethics, and principlism.

We chose principlism as the basis of our advisory algorithm because it provides a set of decision factors common across case types which lends itself to being translated into machine-readable values. Tom Beauchamp and James Childress first proposed principlism in the 1979 edition of their *Principles of Biomedical Ethics*. Famously, their set of prima-facie principles comprises beneficence, non-maleficence, respect for patient autonomy and justice. But how does one build an algorithm around these principles?

**Model and Training**

As the technical solution for implementing Beauchamp and Childress’ prima-facie approach, we chose a type of machine-learning model known as *fuzzy cognitive maps* (FCMs). FCMs are machine-learning models that can simulate dynamic systems, such as human decision-mak ing processes. The relevant components of the process are mapped onto a causal graph which consists of nodes that are linked by causal connections. Nodes represent the entities or concepts to be modelled. In our case, these are parameters of the respective medical case

– for example, whether the patient has reached the age of majority – as well as higher-level concepts, such as the principle of beneficence.

The connections between nodes are weighted, which means that some factors can have a stronger influence on intermediate or output nodes than others. On the basis of the input values of a case and the connections between nodes, the FCM can simulate the causal interactions between the decision-relevant concepts over time.

At each simulation step, every node in the network aggregates all the values of the concepts by which it is influenced, weighted by its incoming causal connections. The node that represents the principle of autonomy, for example, aggregates a number of factors – such as whether the patient has decisional capacity. Each factor is weighted by the strength of the incoming connection between the factor and the autonomy node. Once a node has weighted and aggregated the values of all the nodes by which it is influenced, the resulting number is mapped to a value between 0 and 1 using an S-shaped activation function. Each node then passes its aggregated activation onto all the nodes that are influenced by it – again via weighted connections. The autonomy node may, for example, transmit its value to a node that represents whether to follow the patient’s treatment preference. This process is repeated over several time steps until the system stabilises. The algorithm then reports its recommendation regarding the intervention in question in an output node.

One can either manually specify the strength and the polarity of the connections between the nodes in an FCM or acquire them from input examples through various forms of machine learning. We chose a *genetic algorithm*, which is a machine-learning method in spired by evolutionary biology. Genetic algorithms start out from a pool of random guesses that are refined and improved over time. For each ‘generation’, the algorithm identifies the guess, that is, the set of FCM connection weights, that delivers the best results – namely, those that are closest to the solution provided by human ethicists. The guesses are used to breed a new and even better generation of solutions. At certain intervals, random mutations are introduced to add some variation to the solution pool. This process is repeated until a certain performance threshold is reached or until no further improvement is observed over a set number of generations.

**User Interface**

With the technical architecture in place, the next step was to provide the algorithm with the input categories necessary to capture the specific parameters of individual cases. To do this, we identified the variables that usually play a role in case discussions of clinical ethics committees. The algorithm begins by asking general questions about the patient’s current health status to obtain the background against which any positive or negative consequences that the treatment may yield will be compared. The algorithm then proceeds to request the values for the variables that underlie Beauchamp and Childress’ principles of beneficence, non maleficence, and autonomy. The questions are grouped accordingly.

**Dataset**

Algorithms must be trained before they function properly. For supervised learning, one provides a model with a range of inputs and a corresponding set of predefined solutions, the so-called *labels*. The algorithm is then trained to learn a mapping that will most closely match the inputs in its dataset to the given answers.

When machine intelligence is used, for instance, to detect tumors in images of stained tissue slides taken from biopsies, pathologists label each image of the training data set that is fed into the algorithm regarding whether it contains healthy or abnormal tissue. In the field of ethics, these labels are much more difficult to obtain. Ethically relevant situations rarely admit of ‘objectively correct’ solutions. What is morally right can be highly controversial. However, the supervised training of algorithms requires definite answers – in our case whether to recommend a certain medical intervention or to advise against it. How could one acquire this data?

Cases that were brought before clinical ethics committees emerged as the most suitable data source. While it is important to note that even the decisions that these committees have reached cannot be regarded as morally objective either, they still promise to deliver adequate training data for algorithms whose goal it is to replicate as closely as possible the recommendations that ethical advisory bodies typically issue.

**What Will the Future Bring?**

In this first feasibility study, we propose a way in which machine intelligence could be utilised to solve a range of real-life moral dilemmas that occur in clinical settings. Our algorithm offers a framework to systematically break down medical ethics cases into a set of quantifiable parameters and provides a novel approach to modelling their assessment in a computerised way.

That one *can* do something does not imply that one also *should*: the basic technological means to aid ethical decision-making now exist; but would it really be a good idea to implement such a technology in our clinics? Most people will find the prospect of autonomously driving vehicles taking morally relevant decisions in situations of unavoidable harm easier to accept than having judgments in clinical settings made by machine intelligence.

In the foreseeable future, ethical advisory systems will likely be employed only to support, not to stand in lieu of, human judgment. They could, for example, be used for educational purposes such as for training medical students and aspiring ethicists. One may also utilise them to provide patients and relatives with informal ethical guidance in cases that are not deemed important or controversial enough to be brought before clinical ethics committees.

Currently, the prospect of putting our patients’ fate into the hands of non-biological apparatuses is met with great resistance. Irrespective of whether or not this is a path that society will ultimately wish to pursue – it is crucial already to begin this discussion and carefully to consider the virtues and vices of the novel options that are becoming available to us.