



## Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept

Lukas J. Meier, Alice Hein, Klaus Diepold & Alena Buyx

**To cite this article:** Lukas J. Meier, Alice Hein, Klaus Diepold & Alena Buyx (2022) Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept, The American Journal of Bioethics, 22:7, 4-20, DOI: [10.1080/15265161.2022.2040647](https://doi.org/10.1080/15265161.2022.2040647)

**To link to this article:** <https://doi.org/10.1080/15265161.2022.2040647>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 16 Mar 2022.



Submit your article to this journal [↗](#)



Article views: 7562



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 18 View citing articles [↗](#)

## Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept

Lukas J. Meier<sup>a,b</sup> , Alice Hein<sup>a</sup> , Klaus Diepold<sup>a</sup> , and Alena Buyx<sup>a</sup> 

<sup>a</sup>Technical University of Munich; <sup>b</sup>University of Cambridge

### ABSTRACT

Machine intelligence already helps medical staff with a number of tasks. Ethical decision-making, however, has not been handed over to computers. In this proof-of-concept study, we show how an algorithm based on Beauchamp and Childress' prima-facie principles could be employed to advise on a range of moral dilemma situations that occur in medical institutions. We explain why we chose fuzzy cognitive maps to set up the advisory system and how we utilized machine learning to train it. We report on the difficult task of operationalizing the principles of beneficence, non-maleficence and patient autonomy, and describe how we selected suitable input parameters that we extracted from a training dataset of clinical cases. The first performance results are promising, but an algorithmic approach to ethics also comes with several weaknesses and limitations. Should one really entrust the sensitive domain of clinical ethics to machine intelligence?

### KEYWORDS

Algorithms; artificial intelligence; Beauchamp and Childress; clinical ethics; decision-making; machine learning

## INTRODUCTION



In many areas of medicine, time-consuming and labor-intensive duties are on the brink of being delegated to machines (World Health Organization 2021). Increasingly sophisticated algorithms are being developed to interpret medical images (Badgeley et al. 2019; Esteva et al. 2017; Heijden et al. 2018; Madani et al. 2018; Rajpurkar et al. 2018; Serag et al. 2019; Tong et al. 2020; Wang et al. 2020), analyze electrocardiograms (Hannun et al. 2019), predict long-term therapeutic outcomes (Avati et al. 2018; Komorowski et al. 2018), aid in precision dosing (Angehrn et al. 2020), detect signs of mental disorders (Laacke et al. 2021) and even deliver psychological therapies (Fiske, Henningsen, and Buyx 2019).

Inherently *ethical* tasks, however, have so far been excluded from the promise of automation. As of today, no machine-intelligence systems exist that are designed specifically for the making of sophisticated moral decisions (Cervantes et al. 2020). To the best of our knowledge, the creation of an algorithmic advisory system for clinical ethics has only been attempted once, and it was not developed beyond the early prototype stage. The agent, described in a pioneering article (Anderson, Anderson, and Armen 2006), was

confined to a single type of dilemma situation, namely, one in which mentally competent patients refuse to undergo treatments that would be beneficial for them. The algorithm was then to decide whether medical staff should accept this decision or challenge it. An updated version also included a scenario in which patients are reminded to take their medication (Anderson and Anderson 2018).

In clinical reality, of course, a multitude of different types of ethical dilemmas arise. Possible solutions are often much more controversial than in the case of informed refusal of treatment with full decisional capacity—a patient's right that is enshrined in law. To be of help to medical staff and patients, an algorithm would have to be able efficiently to handle a wide variety of ethical problems. In this proof-of-concept study, we shall present a first attempt at developing such an advisory system.

We will begin by exploring which normative ethical theory could act as the *moral* basis of the algorithm. Next, we shall consider different *technical* approaches and detail why we chose fuzzy cognitive maps to set up our *Medical Ethics Advisor* METHAD. We will explain how we captured the parameters of clinical cases and solved these dilemma situations by

**CONTACT** Lukas J. Meier  [ljm204@cam.ac.uk](mailto:ljm204@cam.ac.uk)  Institute of History and Ethics in Medicine, Technical University of Munich, Ismaninger Strasse 22, München, 81675, Germany.

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

implementing Beauchamp and Childress' prima-facie principles. Finally, we shall evaluate the algorithm's performance, consider challenges and limitations of our approach and reflect on the ethical implications of employing such a technology in the clinic.

## THE ETHICAL BASIS: PRIMA-FACIE PRINCIPLES

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 (Allen, Varner, and Zinser 2000; Gips 2011). 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. The act that produces the best overall result is the one to choose. The most prominent type of consequentialist ethics is utilitarianism (Bentham 1789; Mill 1863; Sidgwick 1877).

Conversely, deontologists hold that actions are morally right when they conform to a particular *norm* or set of norms. Actions are therefore regarded as innately ethical or innately unethical independent of their respective consequences. Many variants of deontological moral systems have been proposed, of which the most influential one is the Kantian (Kant 1993).<sup>1</sup>

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* (Aristotle 1995; Hume 1826). Aretaic ethics originated in ancient Greek philosophy (Aristotle 1995; Plato 1997) and were also influential in the scholastic period (Aquinas 1981).

Given the three theories' complementary strengths and weaknesses—which we do not need to rehash (see Copp 2011)—even hundreds of years of philosophizing 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 despite the fact that the philosophical debate will likely never reach a definitive conclusion. To take an often-discussed example: autonomous vehicles must

be equipped with principles that specify what to do in situations of unavoidable harm. A 2016 study found that the vast majority of people advocate utilitarian decision rules to minimize the overall number of casualties, while at the same time preferring a car that—according to a deontological principle—protects its passengers at all cost (Bonnefon, Shariff, and Rahwan 2016). Such puzzles plague the implementation of machine intelligence in most ethically relevant fields.

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. Moreover, none of the three types of basic ethical theory provide the tools to enable concrete decision-making in actual clinical cases (Flynn 2021). Less general approaches of greater practical applicability were developed—among them casuistry (Jonsen 1991), narrative ethics (Charon and Montello 2002; Montello 2014), feminist ethics (Sherwin 1992; Wolf 1996) and principlism (Beauchamp and Childress 2013).

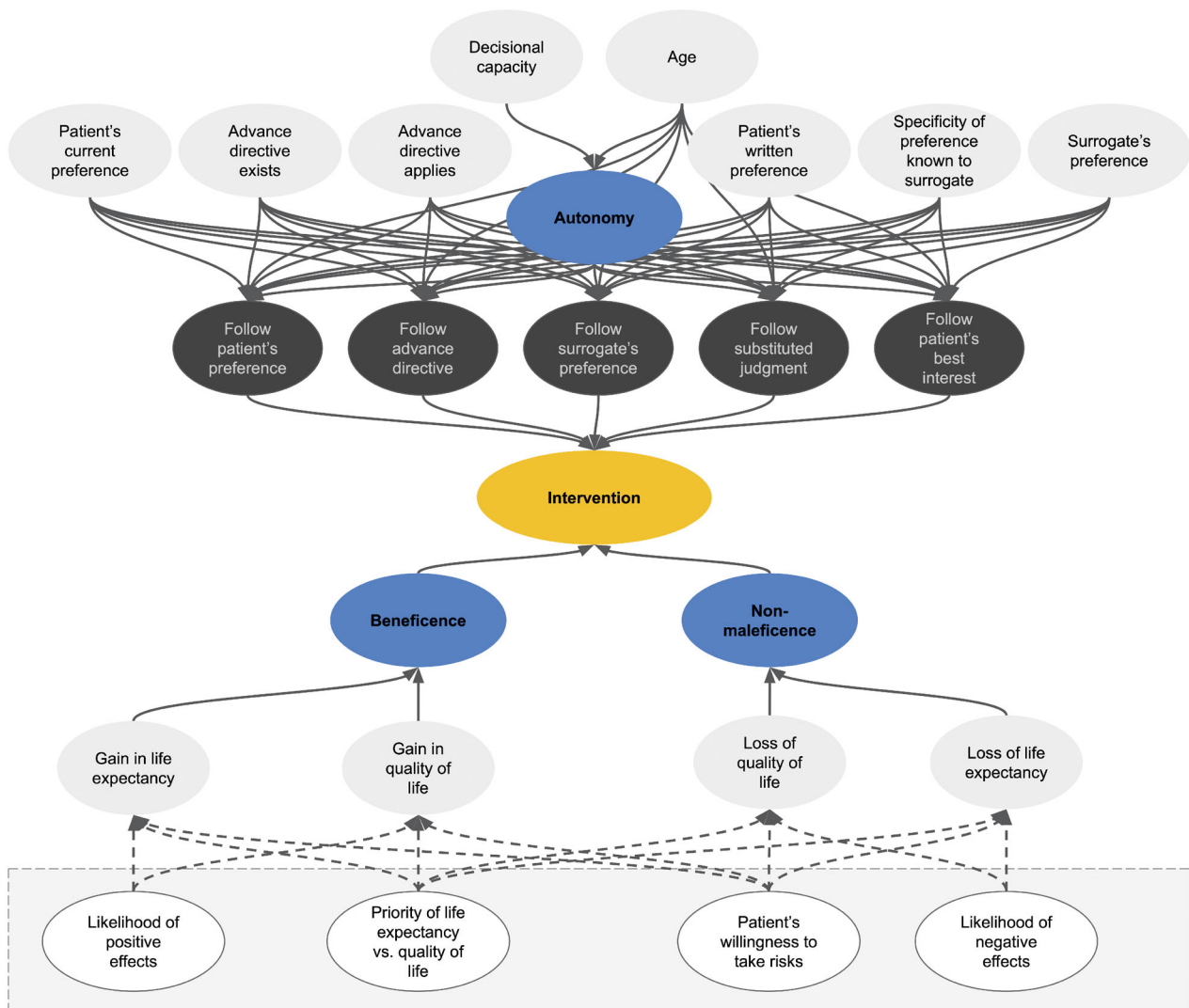
While discussing in detail the strengths and weaknesses of these and other relevant accounts would exceed the scope of this paper (see Flynn 2021), it is certainly fair to say that all have made important contributions to ethical medical decision-making. After much deliberation, we chose principlism 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. Moreover, many authors (Gillon 2015; Veatch 2020), though certainly not all (Tong 2002), regard principlism as the most influential methodology for doing bioethics.

Tom Beauchamp and James Childress first proposed principlism in the 1979 edition of their seminal *Principles of Biomedical Ethics*. Famously, their set of four prima-facie principles comprises beneficence, non-maleficence, respect for patient autonomy and justice (Beauchamp and Childress 2013). Derived from people's everyday moral convictions, this mid-level approach offers, so they argue, a basis for the structured analysis of medical cases in spite of the unresolved fundamental moral disputes. But how does one build an algorithm around Beauchamp and Childress' principles?

## THE TECHNICAL BASIS: FUZZY COGNITIVE MAPS

As the technical solution for implementing Beauchamp and Childress' prima-facie approach, we

<sup>1</sup>For an overview of the single- and multi-rule moral systems that have featured in medical ethics, see Veatch (2020).



**Figure 1.** Visualization of the METHAD FCM. In this pilot study, we omitted the principle of justice for reasons explained below.

chose a type of machine-learning model known as *fuzzy cognitive maps* (FCMs). The term *machine learning* refers to any algorithm that learns from, and improves with, experience. FCMs are machine-learning models that can simulate dynamic systems, such as human decision-making processes. Presently, their areas of application include the aggregation of differing expert opinions in software project management (Pérez-Teruel, Leyva-Vázquez, and Estrada-Sentí 2015), the evaluation of developmental prospects in the industry (Chen and Chiu 2021) and the analysis of risks in clinical drug administration (Mazzuto et al. 2018). 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 modeled. In our study, 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 Beauchamp and Childress' principle of beneficence (Figure 1).

The causal connections between nodes can be positive, negative or neutral. The patient's capacity to consent to a treatment and the node that represents the principle of autonomy, for example, are positively connected since an increase in the former would lead to an increase in the latter. An input like a decrease in the patient's quality of life and the node that stands for the concept of non-maleficence, on the other hand, are negatively connected because lowering the quality of life involves a decrease in non-maleficence. 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 and whether they have reached the age of majority. 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. One does this to keep the values of all nodes—also called their *activation*—in a comparable range. 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 stabilizes (Felix et al. 2019). The algorithm then reports its recommendation regarding the intervention in question in an output node. It takes values between 0 and 1, where 0 means strongly opposed to the intervention, and 1 means strongly in favor of it.

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 inspired 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 given solution (how we obtained these solutions is explained below). 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 (Reeves 2010).

FCMs occupy a space in between two artificial-intelligence paradigms: deep neural networks and symbolic methods. Deep neural networks are a subcategory of machine-learning models. They consist of thousands or even millions of artificial neurons and synapses arranged in multiple layers—hence the name “deep.” These methods have received widespread

attention for achieving impressive results in a broad range of application areas, such as image classification or gameplay, although their robustness in real-world domains has recently been called into question (D’Amour et al. 2020). The way in which deep neural networks and FCMs process their inputs by passing activations on through weighted connections is very similar. However, FCMs are usually much smaller—comprising only 21 nodes and 54 connections in our case—and can be applied to domains with relatively little training data. In contrast, deep neural networks learn from extremely large sets of examples. Since collecting ethically relevant medical cases is a labor-intensive manual task that requires categorization and expert data labeling, it was important that our algorithm be able to learn from comparatively few data points.

Besides the difference in the required amount of training data, an important distinction between FCMs and deep neural networks is the fact that each node in an FCM has a human-designated semantic meaning. It is, for instance, explicit which node in our FCM represents autonomy and which stands for beneficence, whereas deep neural networks decide internally what each of their neurons shall represent. Usually, these designations are not human-interpretable, which is why deep-learning methods are often described as *black boxes* (Braun et al. 2021; Watson et al. 2019). In addition to causing difficulties with regard to trustworthiness and accountability, these methods would therefore not have been conducive to an important goal of our study, which is to achieve transparency regarding the ethical knowledge that METHAD acquires. In contrast, the human-readable visualization as causal graphs that FCMs offer make them well-suited tools for interdisciplinary projects such as this pilot study (Jetter 2006). Their interpretability enabled the medical ethicists on our team to provide feedback on which concepts should be included in the model and which nodes should be causally connected. Our programmers were then able to encode this domain knowledge in a computer-readable way.

Opposite the deep-learning paradigm on the artificial-intelligence spectrum are symbolic methods. These methods deal with logical reasoning or search over *given* high-level knowledge representations. A programmer might, for example, specify a rule according to which *if* a patient is fully capable to consent *and* has reached the age of majority, *then* doctors should adhere to their treatment preferences. Methods like decision trees or integer logic programming—the

approach that Anderson, Anderson, and Armen (2006) pursued in their study—can extract such rules from relatively few examples. Since these rules can be expressed in natural language, they are also highly transparent. However, several problems arise when one applies symbolic methods to a domain like medical ethics. First, symbolic methods operate in the realm of *true* and *false*, whereas case discussions in clinical ethics require much more nuanced options. Thus, methods like integer logic programming struggle when data is not clear-cut (Evans and Grefenstette 2018). Secondly, medical case discussions often comprise various factors that interact in complex ways, so that accounting for all of them would result in a very long list of convoluted rules.

For the task of issuing recommendations in medical ethics, FCMs therefore represent a good compromise between deep learning and symbolic methods. Like symbolic methods, they require relatively little data; but like neural networks, they are able to incorporate vague and more uncertain knowledge through their weights and activations and thereby to establish an adequate balancing of Beauchamp and Childress' principles in a context-dependent manner.

In terms of interpretability, FCMs are located between the black boxes of deep learning and the clear-cut rules of symbolic methods in that their nodes as well as their connections have a clear, interpretable meaning. By inspecting the weights that the network has learned, one can determine whether the strength and the polarity of the connections match ethicists' suggestions. This is in contrast to classical neural networks, whose nodes and connections have no obvious meaning (Felix et al. 2019).

## CAPTURING CASE PARAMETERS

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. We had to tread a fine line: if the interface overburdens the users by demanding of them that they input a multitude of complicated parameters, the advantage of a computerized processing of data is markedly decreased; if, on the other hand, the number of input variables is so small that the peculiarities of the respective cases are not conveyable to the algorithm, the latter can only reach very generalized verdicts and is limited to providing overly

generic advice that would be useless for dealing with the highly unique real-life clinical cases.

We therefore strove for a balance between user-friendliness and precision of outcome and created a catalog of twenty input parameters per case. METHAD 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. Medical staff can either enter the values themselves or jointly with the patient. Patient participation is especially desirable where subjective parameters or personal preferences are at stake. Consequently, we designed the user interface in a way that is also accessible to laypeople.

## Beneficence and Non-Maleficence

The principle of *beneficence* demands of medical personnel that they attend to the patient's welfare (Beauchamp and Childress 2013). For operationalizing this principle, two factors are of primary importance: the intervention's influence on the patient's *life expectancy* and on their *quality of life* (Buyx, Friedrich, and Schöne-Seifert 2009; Ventegodt, Merrick, and Andersen 2003). While the former is a quantifiable value, the latter is an inherently subjective category that must somehow be translated into a machine-readable form.

A patient's quality of life does not only encompass purely physiological parameters but extends also to cognitive, emotional, and social components (Bullinger 2014; Gutmann 2017; Wilm, Leve, and Santos 2014; World Health Organization 2012). Life quality can be assessed with various questionnaires, like the *SF-36 Health Survey*, the *EQ-5D* or the *WHO-QoL Questionnaire* (Kohlmann 2014). Since these tools have different strengths and weaknesses and are usually most reliable when tailored to a specific malady, we refrained from embedding any of the surveys directly into our user interface.

There are often discrepancies between the subjectively perceived quality of life and the patient's clinically assessed health status (Bullinger 2014; Perron, Morabia, and de Torrenté 2002; Woopen 2014). Especially in the case of disabilities, third-personal evaluation frequently underestimates the patient's

**Overview**

- Start
- Patient
- Beneficence
- Non-Maleficence
- Autonomy
- Results

## Beneficence

Please describe the expected *positive* effects of the intervention in question.

You have indicated that - untreated - the patient would probably have 5 years left to live.

How many years of life would he or she probably have left to live if the intervention was begun (or continued)?

0.00 23.00 100.00

You have rated the patient's current quality of life as "poor".

If the intervention has a potential positive effect on the patient's quality of life, how will it be affected?

- no positive effect
- marginally improved
- moderately improved
- markedly improved

For how many years can this positive effect on the quality of life be expected to persist?

0.00 23.00 23.00

What is the likelihood for (the continuation of) the intervention to yield these positive effects?

- very low
- low
- fair
- high
- very high

Figure 2. User interface: beneficence.

perceived quality of life because individuals who have not experienced a particular malady usually fail to adequately conceptualize it (Drummond et al. 2009; Mast 2020). Being able to choose the method of evaluation according to the respective situation is therefore paramount. The aggregated score that the respective questionnaire delivers can then be used as an input parameter.

To what degree an intervention has the potential to improve a particular patient's quality of life or their life expectancy can *ex ante* only be specified probabilistically (Buyx, Friedrich, and Schöne-Seifert 2009). A certain drug may have a great impact on some patients, but be ineffective in others; some patients experience serious side effects, whereas others do not. METHAD therefore asks the user not only to specify the improvement in these two parameters that the intervention is assumed to yield, but also the *likelihood* of this effect occurring (Figure 2).

In general, one cannot simply interpret anticipated gains in life expectancy or in the quality of life in absolute terms: being given an additional month of life may mean only a marginal improvement for one patient, but be an extremely favorable outcome for another; likewise, a minute change in the quality of life may be considered too small to even attempt therapy in one case, but be regarded as a great relief in

another (Deutscher Ethikrat 2011). To reflect this interindividual discrepancy, METHAD collects data on what the patient's presumed trajectory would look like if the measure in question would *not* be performed.

Patients also differ greatly in whether they prefer an extension in the length of their life or an improvement in the quality of the latter (Craig et al. 2018). Since it is not uncommon for certain therapy options to achieve one of these goals only at the expense of the other—dialysis, for instance, prolongs a patient's life but simultaneously diminishes its quality—the user interface requests that, if possible, the patient indicate their preference between these sometimes competing aims. While the interface offers itemized answer options for most questions, we realized the input of this highly subject-specific parameter via a continuous slide switch to allow for more fine-grained values.

Many therapies carry great risks or come with significant side effects which must be taken into account when making ethically relevant decisions. Consequently, the principle of *non-maleficence* demands of medical staff that they do not inflict evil or harm (Beauchamp and Childress 2013). This requires anticipating risks, avoiding or minimizing them as far as possible, and monitoring any harms

that may occur on an ongoing basis during treatment. After all data for the category of beneficence has been provided, METHAD's interface therefore guides the user to the questions associated with this second principle. Analogous to the parameters requested for the previous category, the user is now asked to specify whether the planned intervention has the potential to diminish the patient's life expectancy or their quality of life, and if so, how great these adverse effects would presumably be and what the odds are for them to occur. Finally, one must enter the patient's individual willingness to take this risk.

### Respect for Patient Autonomy

Doctors must always act in accordance with the patient's choices and values (Herring 2016). The next principle to implement was therefore the *respect for patient autonomy* (Beauchamp and Childress 2013). Although legal standards and assessment tools vary between countries, there is general agreement that patients have decisional capacity and can thus exercise their autonomy if they are able "to communicate a choice, to understand the relevant information, to appreciate the medical consequences of the situation, and to reason about treatment choices" (Appelbaum 2007, 1835).<sup>2</sup> While an individual's decisional competence admits of various degrees, practicality dictates that one specify a threshold below which the patient is regarded as incompetent (Beauchamp and Childress 2013). In clinical practice, the judgment about a patient's capacity is therefore usually binary: the patient is declared either capable or incapable to take the decision.

In recent years, a growing number of authors have pointed out that even when patients are deemed to be lacking decisional capacity, they are often still able to express preferences, and argued that the latter should, if possible, be taken into account (Berlinger, Jennings, and Wolf 2013; Jaworska 1999; Navin et al. 2021; Walsh 2020; Wasserman and Navin 2018). In response, we decided to provide the user with the option of passing more nuanced judgment by offering two additional choices—marginally and moderately capable—in between the customary all-or-nothing categories "fully capable" and "not capable." The algorithm would learn to attach more weight to the patients' preferences the higher they score on this scale and depending on whether they have reached

the age of majority.<sup>3</sup> However, conceiving of capacity as a spectrum rather than as a binary construct is not currently common practice and is thus just offered as an experimental addendum.

When the user rates the patient's decisional capacity as impaired, METHAD should follow the standard hierarchy (Bundesärztekammer 2018; Schweizerische Akademie der Medizinischen Wissenschaften 2013) of also taking into account other sources for establishing preferences. The first source to consult is the advance healthcare directive. When an advance directive is valid—in most jurisdictions this means that it was signed after having reached the age of majority, while in possession of full capacity, that it was not revoked and that the document is in accordance with the law—it should normally be followed unconditionally.<sup>4</sup>

It is often the case, however, that patients expressed their wishes only vaguely, or, if they did express them clearly, what they have laid down may not be applicable to the present situation (Fagerlin and Schneider 2004). Initiatives like advance care planning seek to improve this situation in clinical practice (Detering et al. 2010). To accommodate this difficulty, we integrated a parameter into the algorithm with which one can rate how well the written instructions correspond to the decision that must be taken. The algorithm should learn to give more weight to the patient's preferences when their respective applicability is high (Figure 3).

When a patient has lost the capacity to take decisions but had not drafted an advance directive, or when the latter does not apply to the current situation, most legal systems permit that a surrogate decision maker be appointed. Usually, this is the patient's next of kin or an individual nominated by the courts. The surrogate's duty is to make decisions on the patient's behalf according to the substituted judgment standard. Any choice made must reflect the incapacitated person's preferences and values as closely possible (Batteux, Ferguson, and Tunney 2020). To honor this requirement, METHAD inquires about how well the surrogate is informed about the patient's wishes and should adjust the emphasis given to the surrogate's recommendation accordingly.

<sup>2</sup>See also Herring (2016) and Schweizerische Akademie der Medizinischen Wissenschaften (2013).

<sup>3</sup>In some jurisdictions, the right to consent to certain medical interventions is independent of whether the patient has reached the age of majority (Griffith 2016). To employ the algorithm in these areas, one would therefore need to remove this variable.

<sup>4</sup>We leave aside the often-discussed concern that the choices documented in living wills may not be stable across time and, consequently, fail to accurately reflect the individual's later preferences during an incapacitating illness (Beauchamp and Childress 2013; Mast 2020; Wasserman and Navin 2018).



**Autonomy**

Please rate the patient's capacity to consent to the intervention in question. A patient possesses full decisional capacity if he or she is able to understand the relevant information, appreciate the medical consequences of the situation, reason about treatment choices, and communicate his or her choice.

Does the patient possess the capacity to consent to the intervention in question? The patient is...

definitely incapable

Has the patient reached the age of majority?

Yes

What is the patient's current preference regarding the treatment in question?

not known

Did the patient draft an advance directive that complies with the law and is valid? An advance directive is valid if it was signed after having reached the age of majority, with full capacity, and has not been revoked.

Yes

What is the patient's written preference regarding the treatment in question?

definitely not treat

How applicable to the current situation are the wishes that are specified in the advance directive?

fully applicable

What does the surrogate decision maker recommend in the interest of the patient?

no surrogate appointed

Previous Next

Figure 3. User interface: patient autonomy.

Finally, if the patient's wishes are entirely unknown, most legal systems follow the best-interest approach, which specifies that the therapeutic option is to be chosen that—measured by intersubjective standards—promises to yield the best clinical outcome. When this is the case, METHAD is supposed to derive its recommendation solely from considering the values entered for the categories beneficence and non-maleficence.

### Justice

The fourth and final of Beauchamp and Childress' (2013) principles—*justice*—specifies that available health-care resources should be distributed fairly. Does this mean that they shall be allocated in a way that maximizes utility or do certain deontological rules take priority? That the principle of justice does not provide an answer to this question is due to the great plurality of theories of distributive justice. Traditionally, one distinguishes between four major types of theories: the *libertarian*, which place emphasis on individual freedom and autonomous choice (Locke 2003); the *communitarian*, which are concerned with the welfare of the collective (Hegel 1991); the *egalitarian*, which highlight equal access to the available

goods (Rawls 1999); and the *utilitarian*, which strive to maximize well-being for the greatest possible number of individuals (Bentham 1789; Mill 1863). No consensus exists as to which theory should guide the distribution of medical resources in society (Marckmann 2001). Hence, while Beauchamp and Childress' prima-facie principles help to evade conflicts between the major ethical theories at the top level, a somewhat analogous conflict reappears at the level of the principle of justice.

Countries' health policies vary significantly in their adherence to these theories of justice (Beauchamp and Childress 2013; Buyx, Friedrich, and Schöne-Seifert 2009; Deutscher Ethikrat 2011; Drummond et al. 2009). States with universal health-care systems lean more toward egalitarian values, whereas pluralist systems sympathize primarily with libertarian principles and freedom of choice for patients, often to the detriment of communitarian considerations. The British NHS, for instance, takes a cost-effective and benefit-maximizing approach in which quality-adjusted life years (QALYs) play a major role in allocating medical resources, which are made available to everyone (Herring 2016; National Health Service 2021; Paulden 2017). Health care in the US, in contrast, traditionally adheres to libertarian ideals. Citizens must basically

protect their health on their own initiative (Beauchamp and Childress 2013). How widely countries differ in what they deem a just way of distributing medical goods was recently revealed with the publication of their respective guidelines for the allocation of beds in intensive-care units during the COVID-19 pandemic (Ehni, Wiesing, and Ranisch 2021; Joebges and Biller-Andorno 2020; Jöbges et al. 2020; Lewandowski and Schmidt 2020). Finally, even if one could agree on a theory of distributive justice, there would still be many different ways in which fairness could be *calculated* (Verma and Rubin 2018).

Given these great discrepancies in the practical operation of the different health-care systems as well as the lack of consensus regarding the underlying theory of what would constitute a fair distribution of medical resources, incorporating the principle of justice into our algorithm was not achievable without making specific, and possibly unwarranted, health-political and socio-economical background assumptions, of which we wished to stay clear in this pilot study. Consequently, METHAD passes judgment only at the level of the individual patient. Matters of intersubjective justice—for example, whether a certain treatment option would be too costly—must be resolved by the user.

## TRAINING THE ALGORITHM

Algorithms must be trained before they function properly. For supervised learning, which is the most common form of machine 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 dataset that is fed into the algorithm regarding whether it contains healthy or abnormal tissue (Pantanowitz et al. 2020). In the field of ethics, these labels are much more difficult to obtain. Ethically relevant situations rarely admit of “objectively correct” solutions. As we have seen, 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?

We considered three potential sources for compiling training datasets: court judgments, surveys conducted on the general public’s moral preferences, and decisions from clinical ethics committees. We deemed court rulings unfitting for the purpose since several factors play decisive roles in these verdicts that are irrelevant in questions of medical ethics. What is ethical and what is legal is not always congruent (Herring 2016). Surveys of moral preferences (Awad et al. 2018; Uldall 2015) bring with them the difficulty that they cannot possibly reflect the complexity that the making of patient-centered clinical decisions requires. The sheer amount of detail that comes with every single case precludes a broad survey-based approach. In the end, 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, 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.

To mitigate regional influences and national differences, we did not gather historical cases from the ethics committee in our own university hospital, but sourced them from larger collections in the literature (Ackerman and Strong 1989; Dickenson, Huxtable, and Parker 2010; Freeman and McDonnell 2001; Johnston and Bradbury 2016; Pence 2017; Perlin 1992; Snyder and Gauthier 2008). For every included case, we established the respective values of the parameters described in the foregoing section and fed them into the database. Simultaneously, we also entered the training label: whether or not the intervention in question was recommended or rejected. Often the sources already suggested a preferred solution. Where they did not, our institute’s ethics team specified which decision clinical ethics committees would most likely take. Thus, the algorithm gradually learned which constellations of input parameters are supposed to be associated with which ethical outcome. As shown in Table 1, the dataset covers a wide range of case types.

To increase METHAD’s accuracy, we employed data augmentation—a technique whereby slightly modified versions of the original cases are introduced into the dataset (Shorten and Khoshgoftaar 2019). In each of these variations, we made alterations only to a single value while holding the other parameters constant. This enables the algorithm to discern exactly which factor is supposed to exert an influence on the

**Table 1.** Breakdown of dataset by case category.

Case category	Number of cases (absolute)	Number of cases (relative) (%)
Beginning of life, pregnancy and abortion	8	12
Consent in minors	11	16
Advance directives and consent in adults	13	19
Patient's refusal of treatment	11	16
Request or provision of futile treatment	9	13
Withholding or withdrawal of treatment	14	20
Mental health	3	4
Total	69	100

Since no universally accepted taxonomy of clinical ethics cases exists, authors employ very different categorizations (Ackerman and Strong 1989; Freeman and McDonnell 2001; Herring 2016; Johnston and Bradbury 2016; Pence 2017; Perlin 1992; Stauch, Wheat, and Tingle 2018). Consequently, the classifications displayed here are not to be regarded as mutually exclusive; some cases naturally fit into more than one category. Importantly, METHAD is not dependent on case categorizations. We only include these to illustrate the composition of the initial training dataset.

final recommendation and which parameters are less important in which overall constellations.

## EVALUATION

Observing METHAD issuing its first recommendations was fascinating. In the majority of test cases, its suggestions were surprisingly well in accordance with the solutions obtained from the textbooks and from our ethicists (Figures 4 and 5).

However, there was one case category that caused some problems: the competent refusal of medically highly beneficial treatment. In textbooks for medical students, this situation is often illustrated using the example of Jehovah's Witnesses' informed rejection of blood transfusions (Dickenson, Huxtable, and Parker 2010; Johnston and Bradbury 2016). Here, our algorithm's recommendation for treating the fully competent patients against their will was consistently between 0.3 and 0.6 higher than the training label. Although overriding a patient's decision to prevent significant harm can in certain circumstances be morally justified, most ethicists agree that a patient's *substantial* autonomy interests must not be infringed even if this would result in their certain death (Beauchamp and Childress 2013).

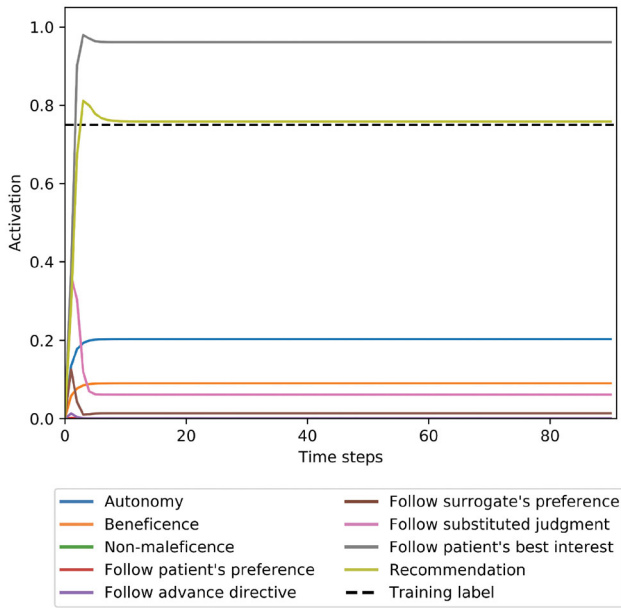
Why did METHAD issue this output? In many of our training cases, a high likelihood of significant gains in the length of life in conjunction with negligible side effects and a good quality of life were associated with a strong tendency to attempt the intervention in question. That patient autonomy can sometimes override this intricate calculus was apparently difficult for the algorithm to pick up. We alleviated this deviation by introducing additional case variations in which autonomy is the deciding factor into the training dataset to reinforce correct behavior in these types of situations (Figure 6).

Currently, the algorithm's database consists of only 69 cases. To evaluate METHAD's performance, we

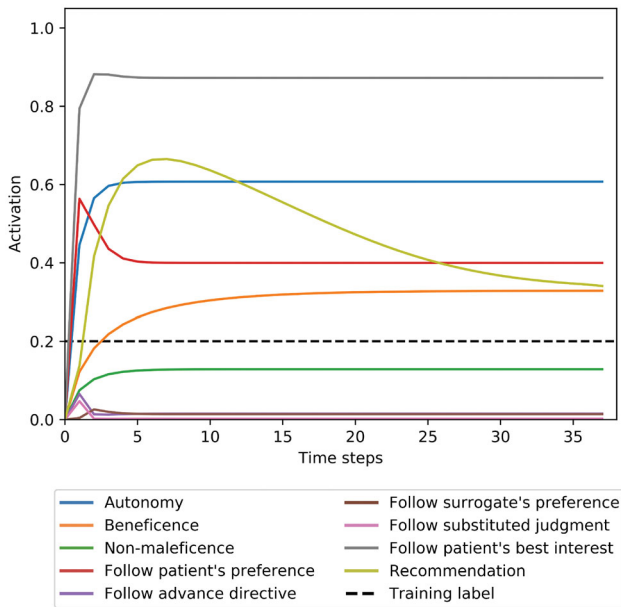
therefore employed stratified k-fold cross-validation, which is a method commonly used in machine learning for assessment in settings with limited data. Splitting the dataset into a training set and a test set according to a given ratio—we used two thirds for training and one third for testing—enables one to assess the model's performance on data that it has never encountered before. K-fold cross-validation means that this split is done *k* times. The data is shuffled, so that the exact makeup of the training and test sets differs slightly each time. Stratification ensures that the data split is done in a way that results in an equal distribution of a certain feature in the training and test sets. We stratified according to case type. Consequently, even though our training set is twice the size of our test set, they both contain the same proportion of cases concerning, for instance, consent in minors.

We used a *k* of 3, which provided us with three separate pairs of training and test datasets. Since there is an element of randomness involved in genetic algorithms, we trained ten models for each train/test-set pair with the configuration shown in Figure 1 to obtain a more robust performance estimate. The results are averaged over these ten models, and we report them separately for the training and the test set. As mentioned, METHAD's recommendations were allowed to take any value between 0 (strongly opposed to the intervention) and 1 (strongly in favor of the intervention). We set a decision threshold of 0.5, which means that outputs  $\geq 0.5$  are counted as approval, and outputs  $< 0.5$  signify that the intervention in question should not be undertaken.

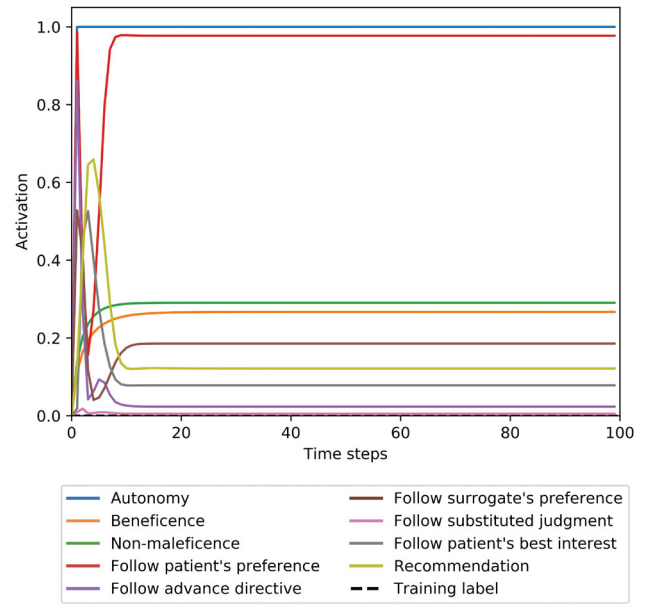
When the algorithm's predictions were compared to the textbook solutions and to our ethicists' judgments, its outputs deviated from these labels on average by 0.11 in the training dataset and 0.23 in the test dataset. METHAD agreed with our ethicists in 92% of the cases in the training set and 75% of the cases in the test set. Note that these metrics measure slightly



**Figure 4.** Example case 1: A 10-year-old is suffering from leukemia. Seeing their child experience the strong side-effects that the therapy induces, the parents want all interventions halted. However, chemotherapy has proven to be highly effective and the child’s prognosis is very promising. METHAD’s analysis indicates that continuing the therapy would very likely be in the young patient’s best interest, which is in accordance with human ethicists’ judgment (denoted as “training label”).



**Figure 5.** Example case 2: A patient with limited decisional capacity requests a medical intervention that is unlikely to result in an extension of her life. There is a strong indication, however, that the treatment would reduce the patient’s quality of life. Her risk preference cannot be established. METHAD begins by analyzing the patient’s wishes, but, after factoring in the intervention’s medical futility, ultimately recommends against proceeding.



**Figure 6.** Example case 3: A competent patient refuses a life-saving treatment. METHAD calculates that the medical benefits for the patient would be enormous – visible as an initial spike – and that the intervention would come with only minimal risks, but, after figuring in that the refusal occurred with full decisional capacity, eventually recommends refraining from intervening.

different things: a prediction may deviate only marginally from the training label but be counted as *incorrect*—for example, if the algorithm’s output is 0.45 and the ethicists’ judgment is 0.6. However, the opposite is also true: a recommendation may be counted as *correct* although the deviation from the label is relatively high—for example, if the output is 0.55 and the label is 1.0 (Table 2).

Although, considering the size of its training dataset, METHAD’s performance is already rather promising, it would be desirable to broaden the database considerably. To further refine the recommendations that the algorithm issues and to make them more robust, hundreds of novel training cases would have to be added. This, however, was not the aim of this first proof-of-concept study. The small example dataset served mostly as a vehicle to test whether the algorithm is, in principle, able to generate reasonable recommendations based on a decision process that is learned from examples.

### LIMITATIONS AND CHALLENGES

We shall now address crucial limitations of our system and consider some challenges that it raises. For the reasons already detailed, we omitted Beauchamp and Childress’ principle of justice in this feasibility study. While *distributive* justice is indeed of little

**Table 2.** Cross-validation results.

Case category	Number of cases	Mean absolute error		Mean binary accuracy	
		Train	Test	Train	Test
Beginning of life, pregnancy and abortion	8	0.10 ± 0.06	0.25 ± 0.14	0.82 ± 0.08	0.68 ± 0.28
Consent in minors	11	0.11 ± 0.07	0.22 ± 0.10	0.94 ± 0.10	0.81 ± 0.12
Advance directives and consent in adults	13	0.10 ± 0.06	0.23 ± 0.10	0.97 ± 0.07	0.75 ± 0.19
Patient's refusal of treatment	11	0.11 ± 0.06	0.18 ± 0.11	1.00 ± 0.01	0.92 ± 0.15
Request or provision of futile treatment	9	0.12 ± 0.07	0.32 ± 0.14	0.87 ± 0.17	0.60 ± 0.27
Withholding or withdrawal of treatment	14	0.09 ± 0.07	0.17 ± 0.11	0.95 ± 0.11	0.83 ± 0.19
Mental health	3	0.18 ± 0.14	0.30 ± 0.16	0.68 ± 0.42	0.27 ± 0.33
Total	69	0.11 ± 0.07	0.23 ± 0.12	0.92 ± 0.10	0.75 ± 0.20

A discrepancy in performance between training and test datasets is to be expected in most machine-learning settings – especially when working with small datasets.

relevance in most individual case discussions, there is another dimension to justice to which one must pay close attention when developing algorithms: *procedural justice*.

When it comes to justice in decision-making, the focus is usually on whether the *outcome* is fair. Studies have shown, however, that people perceive the fairness of the *process* by which a certain decision is reached as equally important—irrespective of whether they agree with the result (Blader and Tyler 2003; Thibaut and Walker 1975). Fairness in the practice of reaching decisions is known as procedural justice.

While authors differ in which elements they take procedural justice to encompass (Leventhal 1980; Solum 2004; Thibaut and Walker 1975), this question now also pertains to the emerging field of algorithmic decision-making. Recent work has been exploring what procedural justice amounts to in this specific context (Lee et al. 2019). The main challenge in setting up algorithms, especially in the health-care sector, is to ensure that they neither amplify existing biases nor introduce novel ones (Char, Shah, and Magnus 2018; Vollmer et al. 2020). Mehrabi et al. (2019) uncover no less than 23 different types of potential biases.

For advisory algorithms like METHAD, this means that utmost care must be exercised in populating the database. Above all, one must collect the chosen cases in a way that is representative and inclusive to ensure that machine intelligence will not make discriminatory decisions (World Health Organization 2021). Given that some parameters of clinical cases are open to interpretation, this requirement also extends to the individuals who enter the data.

Besides the absence of systematic errors in the dataset, there are two main goals in honoring procedural justice in algorithmic decision-making: *transparency* (Braun et al. 2021; Char, Abramoff, and Feudtner 2020) and *control* (Lee et al. 2019).

*Transparency* is crucial to enable all parties involved to retrace how exactly the decision came about (World Health Organization 2021). As described earlier, this requirement led us to exclude machine-learning models with limited explanatory power, such as deep neural networks and support vector machines, which are often criticized for their opacity (Watson et al. 2019). Instead, we chose FCMs, which can be represented as causal graphs with nodes that symbolize the different elements and causal connections of the decision-making process (Figure 1). As shown in Figures 4–6, the procedure of weighing the ethical principles can be visualized as a diagram for every single case. Moreover, METHAD delivers its outputs not in a binary form (intervention recommended/not recommended), but as a digit that demonstrates *how much* in favor or against a certain intervention the algorithm's suggestion is, thereby signposting borderline cases in which greater scrutiny may be needed. Nonetheless, this decision-making process will inevitably be more opaque than well-conducted discussions among skilled clinical ethicists.

Procedural justice also requires *control*: to allow the users to modify as much of the input data as possible and ideally even permit them to influence the inner workings of the algorithm itself (Lee et al. 2019). As described, our analysis of clinical case discussions revealed 20 decisive parameters that appeared repeatedly, and we implemented these into METHAD's user interface. However, only larger clinical trials will show whether special circumstances may require additional input categories. There is always the danger of the so-called *omitted-variable bias*: of failing to take into consideration a parameter that exerts an influence on the result, and thus falsely attributing a certain outcome to the effect only of those variables that were included in the dataset (Mehrabi et al. 2019). If, for instance, the algorithm's user interface did not provide the option to specify a possible decrease in the

patient's *quality* of life caused by the intervention in question, the system would learn to recommend that therapies go ahead whenever the gain in life *expectancy* is great enough (and patient consent is obtained, of course). This would in some cases lead to unethical proposals—for example, when patients could gain an additional year of lifetime only by being subjected to excruciating pain, or when they indicate that they prefer a high quality of life over any temporal prolongation. While machine intelligence will inevitably be inferior to human ethicists in that it will, at least in the beginning, not be able to take into consideration more exceptional peculiarities, algorithmic solutions must still endeavor not to disregard crucial parameters that arise in clinical consultations. Conversely, however, will analyzing which results METHAD yields on the basis of which variables hopefully also help us better to understand the role that each parameter plays in case discussions conducted by human experts.

Regarding control over of the inner mechanics of the algorithm, we have already described how our FCM permits the addition, deletion, and modification of nodes and connections. This way, METHAD can also be adapted to regional, cultural, and juridical differences.

Despite these measures, no algorithm will be fully free from biases, thoroughly transparent, and entirely controllable. While one must endeavor to hold algorithmic decision-making to the same high standards that guide clinicians (Char, Shah, and Magnus 2018), it is therefore essential that mechanisms be in place to override or to appeal decisions (World Health Organization 2021). In the case of METHAD, no autonomous unsupervised use is intended in the foreseeable future.

In a sense, the choice of Beauchamp and Childress' principlism as the ethical basis of our algorithm also constitutes a limitation. Although some authors hold that this approach offers a set of moral commitments "to which all doctors can subscribe, whatever their culture, religion (or lack of religion), philosophy or life stance" (Gillon 2015, 115), others maintain that "morality is more than the attempt either to follow a set of rules or to apply principles appropriately" (Tong 2002, 418). There are situations and particular categories of problems in which casuistic (Jonsen 1991), narrative (Charon and Montello 2002; Montello 2014), feminist (Sherwin 1992; Wolf 1996) or other approaches will be more appropriate. While a synthesis of the different methods and perspectives would certainly be desirable (McCarthy 2003), constructing such an algorithm is not yet feasible. Consequently,

METHAD should only be employed in contexts where one also deems appropriate guidance from Beauchamp and Childress' prima-facie principles.

Lastly, it is worth repeating that the advice that the model offers is about whether medical interventions should be carried out (or treatment continued) or not attempted (or treatment withdrawn). This limits METHAD's applicability to cases in which *treatments* are at issue and excludes the many other areas in which people consult clinical ethicists.

## SHOULD WE EMPLOY IT?

In this first feasibility study, we have proposed a way in which machine intelligence could be utilized to solve a range of real-life moral dilemmas that occur in clinical settings. METHAD offers a framework to systematically break down medical ethics cases into a set of quantifiable parameters and provides a novel approach to modeling their assessment in a computerized way.

We began by choosing the underlying moral theory and analyzing the different technical solutions that are available. We explained why fuzzy cognitive maps are particularly suited for building an ethical advisory system and showed how one can capture the parameters of individual cases. Finally, we reported on the algorithm's performance and considered the limitations of our approach.

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.

Human contact, the intimate relationship between patient and medical personnel, is inherent to the field of medicine as an essential part of making diagnoses, establishing patient preferences and—ultimately—curing (Char, Shah, and Magnus 2018). Could we ever relinquish this bond when it comes to ethical decision-making? Besides medical expertise, should not empathy guide us rather than "cold" computerized calculations?

In the foreseeable future, ethical advisory systems will likely be employed only to support, not to stand in lieu of, human judgment. METHAD could, for example, be used for educational purposes such as training medical students and aspiring ethicists. One

may also utilize it 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.

A time may come, however, at which machine intelligence has become efficient, accurate, and transparent enough to in fact *replace* human ethical decision-making in certain settings. There will be much to be gained from, for example, employing advisory algorithms in overwhelming emergency situations where greater numbers of morally relevant decisions must be taken than would be humanly possible; and, conversely, there will be scenarios in which machine intelligence will likely always remain inferior to human decision-making. Currently, the prospect of putting our patients' fate into the hands of non-biological apparatuses is met with great resistance (Bundesärztekammer 2020). 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.

## ACKNOWLEDGMENTS

We would like to thank Martin Gottwald, Kathrin Knochel, and three anonymous referees.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

## FUNDING

This work was funded by the Technical University of Munich—Institute for Ethics in Artificial Intelligence (IEAI).

## ORCID

Lukas J. Meier  <http://orcid.org/0000-0002-3316-3928>  
 Alice Hein  <http://orcid.org/0000-0002-9457-8131>  
 Klaus Diepold  <http://orcid.org/0000-0003-0439-7511>  
 Alena Buyx  <http://orcid.org/0000-0002-5726-7633>

## REFERENCES

- Ackerman, T. F., and C. Strong. 1989. *A casebook of medical ethics*. New York: Oxford University Press.
- Allen, C., G. Varner, and J. Zinser. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3):251–61. doi: 10.1080/09528130050111428.
- Anderson, M., and S. L. Anderson. 2018. GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics* 9 (1):337–57. doi: 10.1515/pjbr-2018-0024.
- Anderson, M., S. L. Anderson, and C. Armen. 2006. An approach to computing ethics. *IEEE Intelligent Systems* 21 (4):56–63. doi: 10.1109/MIS.2006.64.
- Angehrn, Z., L. Haldna, A. S. Zandvliet, E. Gil Berglund, J. Zeeuw, B. Amzal, S. Y. A. Cheung, T. M. Polasek, M. Pfister, T. Kerbusch, et al. 2020. Artificial intelligence and machine learning applied at the point of care. *Frontiers in Pharmacology* 11 (759):759–12. doi: 10.3389/fphar.2020.00759.
- Appelbaum, P. S. 2007. Assessment of patients' competence to consent to treatment. *The New England Journal of Medicine* 357 (18):1834–40. doi: 10.1056/NEJMc074045.
- Aquinas, T. 1981. *Summa theologiae*. Trans. Fathers of the English Dominican Province. Westminster: Christian Classics.
- Aristotle. 1995. *The complete works of Aristotle*. Edited by J. Barnes. Princeton: Princeton University Press.
- Avati, A., K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah. 2018. Improving palliative care with deep learning. *BMC Medical Informatics and Decision Making* 18 (Suppl. 4):55–64. doi: 10.1186/s12911-018-0677-8.
- Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. 2018. The moral machine experiment. *Nature* 563 (7729):59–64. doi: 10.1038/s41586-018-0637-6.
- Badgeley, M. A., J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley. 2019. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine* 2 (31): 1–10. doi: 10.1038/s41746-019-0105-1.
- Batteux, E., E. Ferguson, and R. J. Tunney. 2020. A mixed methods investigation of end-of-life surrogate decisions among older adults. *BMC Palliative Care* 19 (1):44–12. doi: 10.1186/s12904-020-00553-w.
- Beauchamp, T. L., and J. F. Childress. 2013. *Principles of biomedical ethics*. 7th ed. New York: Oxford University Press.
- Bentham, J. 1789. *An introduction to the principles of morals and legislation*. London: Payne and Son.
- Berlinger, N., B. Jennings, and S. M. Wolf. 2013. *The Hastings Center guidelines for decisions on life-sustaining treatment and care near the end of life*. 2nd ed. New York: Oxford University Press.
- Blader, S. L., and T. R. Tyler. 2003. A four-component model of procedural justice: Defining the meaning of a “fair” process. *Personality & Social Psychology Bulletin* 29 (6):747–58. doi: 10.1177/0146167203029006007.
- Bonnefon, J.-F., A. Shariff, and I. Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352 (6293): 1573–6. doi: 10.1126/science.aaf2654.
- Braun, M., P. Hummel, S. Beck, and P. Dabrock. 2021. Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics* 47: 1–8. doi: 10.1136/medethics-2019-105860.
- Bullinger, M. 2014. The concept of quality of life in medicine: Its history and current relevance. *Zeitschrift für*

- Evidenz, Fortbildung und Qualität im Gesundheitswesen 108 (2–3):97–103. doi: [10.1016/j.zefq.2014.02.006](https://doi.org/10.1016/j.zefq.2014.02.006).t.
- Bundesärztekammer. 2018. Hinweise und Empfehlungen zum Umgang mit Vorsorgevollmachten und Patientenverfügungen im ärztlichen Alltag. *Deutsches Ärzteblatt* 115 (51–52):A2434–A2441.
- Bundesärztekammer. 2020. Orientierungshilfe der Bundesärztekammer zur Allokation medizinischer Ressourcen am Beispiel der SARS-CoV-2-Pandemie im Falle eines Kapazitätsmangels. *Deutsches Ärzteblatt* 117 (20):A1084–A1087.
- Buyx, A. M., D. R. Friedrich, and B. Schöne-Seifert. 2009. Marginale Wirksamkeit als Rationierungskriterium – Begriffsklärungen und ethisch relevante Vorüberlegungen. In *Priorisierung in der Medizin: Interdisziplinäre Forschungsansätze*, ed. W. A. Wohlgenuth and M. H. Freitag, 201–217. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.
- Cervantes, J.-A., S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos. 2020. Artificial moral agents: A survey of the current status. *Science and Engineering Ethics* 26 (2):501–32. doi: [10.1007/s11948-019-00151-x](https://doi.org/10.1007/s11948-019-00151-x).
- Charon, R., and M. Montello, ed. 2002. *Stories matter: The role of narrative in medical ethics*. New York: Routledge.
- Char, D. S., M. D. Abràmoff, and C. Feudtner. 2020. Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics* 20 (11):7–17. doi: [10.1080/15265161.2020.1819469](https://doi.org/10.1080/15265161.2020.1819469).
- Char, D. S., N. H. Shah, and D. Magnus. 2018. Implementing machine learning in health care – addressing ethical challenges. *The New England Journal of Medicine* 378 (11):981–3. doi: [10.1056/NEJMp1714229](https://doi.org/10.1056/NEJMp1714229).
- Chen, C.-T., and Y.-T. Chiu. 2021. A study of dynamic fuzzy cognitive map model with group consensus based on linguistic variables. *Technological Forecasting and Social Change* 171:1–13. doi: [10.1016/j.techfore.2021.120948](https://doi.org/10.1016/j.techfore.2021.120948).
- Copp, D., ed. 2011. *The Oxford handbook of ethical theory*. Oxford: Oxford University Press.
- Craig, B. M., K. Rand, H. Bailey, and P. F. M. Stalmeier. 2018. Quality-adjusted life-years without constant proportionality. *Value in Health* 21 (9):1124–31. doi: [10.1016/j.jval.2018.02.004](https://doi.org/10.1016/j.jval.2018.02.004).
- D’Amour, A., K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv* (arXiv: 2011.03395). <https://arxiv.org/abs/2011.03395>
- Detering, K. M., A. D. Hancock, M. C. Reade, and W. Silvester. 2010. The impact of advance care planning on end of life care in elderly patients: Randomised controlled trial. *British Medical Journal* 340:1–9. doi: [10.1136/bmj.c1345](https://doi.org/10.1136/bmj.c1345).
- Deutscher Ethikrat. 2011. *Nutzen und Kosten im Gesundheitswesen: Zur normativen Funktion ihrer Bewertung*. Berlin: Deutscher Ethikrat.
- Dickenson, D., R. Huxtable, and M. Parker. 2010. *The Cambridge medical ethics workbook*. 2nd ed. Cambridge: Cambridge University Press.
- Drummond, M., D. Brixner, M. Gold, P. Kind, A. McGuire, and E. Nord. 2009. Toward a consensus on the QALY. *Value in Health* 12 (Suppl. 1):S31–S35. doi: [10.1111/j.1524-4733.2009.00522.x](https://doi.org/10.1111/j.1524-4733.2009.00522.x).
- Ehni, H.-J., U. Wiesing, and R. Ranisch. 2021. Saving the most lives—A comparison of European triage guidelines in the context of the COVID-19 pandemic. *Bioethics* 35 (2):125–34. doi: [10.1111/bioe.12836](https://doi.org/10.1111/bioe.12836).
- Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542 (7639):115–8. doi: [10.1038/nature21056](https://doi.org/10.1038/nature21056).
- Evans, R., and E. Grefenstette. 2018. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research* 61 (1):1–64. doi: [10.1613/jair.5714](https://doi.org/10.1613/jair.5714).
- Fagerlin, A., and C. E. Schneider. 2004. Enough: The failure of the living will. *The Hastings Center Report* 34 (2): 30–42. doi: [10.2307/3527683](https://doi.org/10.2307/3527683).
- Felix, G., G. Nápoles, R. Falcon, W. Froelich, K. Vanhoof, and R. Bello. 2019. A review on methods and software for fuzzy cognitive maps. *Artificial Intelligence Review* 52 (3):1707–37. doi: [10.1007/s10462-017-9575-1](https://doi.org/10.1007/s10462-017-9575-1).
- Fiske, A., P. Henningsen, and A. Buyx. 2019. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research* 21 (5):e13216. doi: [10.2196/13216](https://doi.org/10.2196/13216).
- Flynn, J. 2021. Theory and bioethics. In *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta, Spring 2021. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2021/entries/theory-bioethics>.
- Freeman, J. M., and K. McDonnell. 2001. *Tough decisions: Cases in medical ethics*. 2nd ed. Oxford: Oxford University Press.
- Gillon, R. 2015. Defending the four principles approach as a good basis for good medical practice and therefore for good medical ethics. *Journal of Medical Ethics* 41 (1): 111–6. doi: [10.1136/medethics-2014-102282](https://doi.org/10.1136/medethics-2014-102282).
- Gips, J. 2011. Towards the ethical robot. In *Machine ethics*, ed. M. Anderson and S. L. Anderson, 244–253. Cambridge: Cambridge University Press.
- Griffith, R. 2016. What is Gillick competence? *Human Vaccines & Immunotherapeutics* 12 (1):244–7. doi: [10.1080/21645515.2015.1091548](https://doi.org/10.1080/21645515.2015.1091548).
- Gutmann, T. 2017. Gesundheitsbezogene Lebensqualität: Ethische und rechtliche Aspekte. *Frankfurter Forum* 15: 6–13.
- Hannun, A. Y., P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 25 (1):65–9. doi: [10.1038/s41591-019-0359-9](https://doi.org/10.1038/s41591-019-0359-9).
- Hegel, G. W. F. 1991. *Elements of the philosophy of right*. Trans. H. B. Nisbet, ed. A. W. Wood. Cambridge: Cambridge University Press.
- van der Heijden, A. A., M. D. Abramoff, F. Verbraak, M. V. van Hecke, A. Liem, and G. Nijpels. 2018. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn diabetes care system. *Acta Ophthalmologica* 96 (1):63–8. doi: [10.1111/aos.13613](https://doi.org/10.1111/aos.13613).
- Herring, J. 2016. *Medical law and ethics*. 6th ed. Oxford: Oxford University Press.



- Hume, D. 1826. An inquiry concerning the principles of morals. Vol. 4. In *The philosophical works of David Hume*, ed. D. Hume, 235–365. Edinburgh: Black & Tait.
- Jaworska, A. 1999. Respecting the margins of agency: Alzheimer's patients and the capacity to value. *Philosophy & Public Affairs* 28 (2):105–38. doi: [10.1111/j.1088-4963.1999.00105.x](https://doi.org/10.1111/j.1088-4963.1999.00105.x).
- Jetter, A. 2006. Fuzzy cognitive maps for engineering and technology management: What works in practice? Vol. 2. In *Technology management for the global future – PICMET 2006 conference*, 498–512. Istanbul: IEEE.
- Joebges, S., and N. Biller-Andorno. 2020. Ethics guidelines on COVID-19 triage – an emerging international consensus. *Critical Care* 24 (1):1–5. doi: [10.1186/s13054-020-02927-1](https://doi.org/10.1186/s13054-020-02927-1).
- Jöbges, S., R. Vinay, V. A. Luyckx, and N. Biller-Andorno. 2020. Recommendations on COVID-19 triage: International comparison and ethical analysis. *Bioethics* 34 (9):948–59. doi: [10.1111/bioe.12805](https://doi.org/10.1111/bioe.12805).
- Johnston, C., and P. Bradbury. 2016. *100 cases in clinical ethics and law*. 2nd ed. Boca Raton: CRC Press.
- Jonsen, A. R. 1991. Casuistry as methodology in clinical ethics. *Theoretical Medicine* 12 (4):295–307. doi: [10.1007/BF00489890](https://doi.org/10.1007/BF00489890).
- Kant, I. 1993. *Grounding for the metaphysics of morals*. Trans. J. W. Ellington. 3rd ed. Indianapolis: Hackett.
- Kohlmann, T. 2014. Messung von Lebensqualität: So einfach wie möglich, so differenziert wie nötig. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 108 (2–3):104–10. doi: [10.1016/j.zefq.2014.03.015](https://doi.org/10.1016/j.zefq.2014.03.015).
- Komorowski, M., L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. 2018. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 24 (11):1716–20. doi: [10.1038/s41591-018-0213-5](https://doi.org/10.1038/s41591-018-0213-5).
- Laacke, S., R. Mueller, G. Schomerus, and S. Salloch. 2021. Artificial intelligence, social media and depression. A new concept of health-related digital autonomy. *The American Journal of Bioethics* 21 (7):4–20. doi: [10.1080/15265161.2020.1863515](https://doi.org/10.1080/15265161.2020.1863515).
- Lee, M. K., A. Jain, H. J. Cha, S. Ojha, and D. Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW):1–26. doi: [10.1145/3359284](https://doi.org/10.1145/3359284).
- Leventhal, G. S. 1980. What should be done with equity theory? New approaches to the study of fairness in social relationships. In *Social exchange: Advances in theory and research*, ed. K. J. Gergen, M. S. Greenberg, and R. H. Willis, 27–55. New York: Plenum Press.
- Lewandowski, K., and K. W. Schmidt. 2020. Beatmung, Triage und Scoring – Anmerkungen zur Situation in Europa zu Beginn der COVID-19-Pandemie. In *Medizin und Ethik in Zeiten von Corona*, ed. M. Woesler and H.-M. Sass, 35–52. Berlin: LIT.
- Locke, J. 2003. *Two treatises of government and A letter concerning toleration*. Edited by I. Shapiro. New Haven: Yale University Press.
- Madani, A., R. Arnaout, M. Mofrad, and R. Arnaout. 2018. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digital Medicine* 1 (1):1–8. doi: [10.1038/s41746-017-0013-1](https://doi.org/10.1038/s41746-017-0013-1).
- Marckmann, G. 2001. The Eurotransplant kidney allocation algorithm – moral consensus or pragmatic compromise? *Analyse & Kritik* 23 (2):271–9. doi: [10.1515/auk-2001-0209](https://doi.org/10.1515/auk-2001-0209).
- Mast, L. 2020. Against autonomy: How proposed solutions to the problems of living wills forgot its underlying principle. *Bioethics* 34 (3):264–71. doi: [10.1111/bioe.12665](https://doi.org/10.1111/bioe.12665).
- Mazzuto, G., M. Bevilacqua, C. Stylios, and V. Georgopoulos. 2018. Aggregate experts knowledge in fuzzy cognitive maps. *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. Piscataway, NJ: IEEE.
- McCarthy, J. 2003. Principlism or narrative ethics: Must we choose between them? *Medical Humanities* 29 (2):65–71. doi: [10.1136/mh.29.2.65](https://doi.org/10.1136/mh.29.2.65).
- Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv* (arXiv: 1908.09635). <https://arxiv.org/abs/1908.09635>. doi: [10.1145/3457607](https://doi.org/10.1145/3457607).
- Mill, J. S. 1863. *Utilitarianism*. London: Parker, Son, and Bourn.
- Montello, M. 2014. Narrative ethics. *The Hastings Center Report* 44 (Suppl. 1):S2–S6. doi: [10.1002/hast.260](https://doi.org/10.1002/hast.260).
- National Health Service. 2021. *The NHS constitution for England*. London: Department of Health.
- Navin, M., J. A. Wasserman, D. Stahl, and T. Tomlinson. 2021. The capacity to designate a surrogate is distinct from decisional capacity: Normative and empirical considerations. *Journal of Medical Ethics* 1–4. doi: [10.1136/medethics-2020-107078](https://doi.org/10.1136/medethics-2020-107078).
- Pantanowitz, L., G. M. Quiroga-Garza, L. Bien, R. Heled, D. Laifenfeld, C. Linhart, J. Sandbank, A. A. Shach, V. Shalev, M. Vecsler, et al. 2020. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: A blinded clinical validation and deployment study. *The Lancet Digital Health* 2 (8):e407–e416. doi: [10.1016/S2589-7500\(20\)30159-X](https://doi.org/10.1016/S2589-7500(20)30159-X).
- Paulden, M. 2017. Recent amendments to NICE's value-based assessment of health technologies: implicitly inequitable? *Expert Review of Pharmacoeconomics & Outcomes Research* 17 (3):239–42. doi: [10.1080/14737167.2017.1330152](https://doi.org/10.1080/14737167.2017.1330152).
- Pence, G. E. 2017. *Medical ethics: Accounts of ground-breaking cases*. 8th ed. New York: McGraw-Hill.
- Pérez-Teruel, K., M. Leyva-Vázquez, and V. Estrada-Sentí. 2015. Mental models consensus process using fuzzy cognitive maps and computing with words. *Ingeniería Y Universidad* 19 (1):173–88. doi: [10.111144/Javeriana.iyu19-1.mmcp](https://doi.org/10.111144/Javeriana.iyu19-1.mmcp).
- Perlin, T. M. 1992. *Clinical medical ethics: Cases in practice*. Boston: Little, Brown, and Company.
- Perron, N. J., A. Morabia, and A. de Torrenté. 2002. Quality of life of do-not-resuscitate (DNR) patients: How good are physicians in assessing DNR patients' quality of life? *Swiss Medical Weekly* 132 (39–40):562–5. doi: [10.4414/smww.2002.10083](https://doi.org/10.4414/smww.2002.10083).
- Plato. 1997. *Complete works*. Edited by J. M. Cooper and D. S. Hutchinson. Indianapolis: Hackett.
- Rajpurkar, P., J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, et al. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to

- practicing radiologists. *PLOS Medicine* 15 (11):1–17. doi: [10.1371/journal.pmed.1002686](https://doi.org/10.1371/journal.pmed.1002686).
- Rawls, J. 1999. *A theory of justice. Revised Edition*. Cambridge, MA: Harvard University Press.
- Reeves, C. R. 2010. Genetic algorithms. In *Handbook of metaheuristics*, ed. M. Gendreau and J.-Y. Potvin, 2nd ed., 109–139. Boston: Springer.
- Schweizerische Akademie der Medizinischen Wissenschaften. 2013. *Rechtliche Grundlagen im medizinischen Alltag: Ein Leitfaden für die Praxis*. 2nd ed. Basel: Schweizerische Akademie der Medizinischen Wissenschaften.
- Serag, A., A. Ion-Margineanu, H. Qureshi, R. McMillan, M.-J. Saint Martin, J. Diamond, P. O'Reilly, and P. Hamilton. 2019. Translational AI and deep learning in diagnostic pathology. *Frontiers in Medicine* 6 (185): 185–15. doi: [10.3389/fmed.2019.00185](https://doi.org/10.3389/fmed.2019.00185).
- Sherwin, S. 1992. *No longer patient: Feminist ethics and health care*. Philadelphia: Temple University Press.
- Shorten, C., and T. M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data* 6 (1):1–48. doi: [10.1186/s40537-019-0197-0](https://doi.org/10.1186/s40537-019-0197-0).
- Sidgwick, H. 1877. *The methods of ethics*. 2nd ed. London: Macmillan.
- Snyder, J. E., and C. C. Gauthier. 2008. *Evidence-based medical ethics: Cases for practice-based learning*. Totowa: Humana Press.
- Solum, L. B. 2004. Procedural justice. *Southern California Law Review* 78 (181):181–322.
- Stauch, M., K. Wheat, and J. Tingle. 2018. *Text, cases and materials on medical law and ethics*. 4th ed. London: Routledge.
- Thibaut, J. W., and L. Walker. 1975. *Procedural justice: A psychological analysis*. Hillsdale: Erlbaum Associates.
- Tong, R. 2002. Teaching bioethics in the new millennium: Holding theories accountable to actual practices and real people. *The Journal of Medicine and Philosophy* 27 (4): 417–32. doi: [10.1076/jmep.27.4.417.8609](https://doi.org/10.1076/jmep.27.4.417.8609).
- Tong, Y., W. Lu, Y. Yu, and Y. Shen. 2020. Application of machine learning in ophthalmic imaging modalities. *Eye and Vision* 7 (22): 1–15. doi: [10.1186/s40662-020-00183-6](https://doi.org/10.1186/s40662-020-00183-6).
- Uldall, S. W. 2015. How Danes evaluate moral claims related to abortion: A questionnaire survey. *Journal of Medical Ethics* 41 (7):570–2. doi: [10.1136/medethics-2014-102102](https://doi.org/10.1136/medethics-2014-102102).
- Veatch, R. M. 2020. Reconciling lists of principles in bioethics. *The Journal of Medicine and Philosophy* 45 (4–5): 540–59. doi: [10.1093/jmp/jhaa017](https://doi.org/10.1093/jmp/jhaa017).
- Ventegodt, S., J. Merrick, and N. J. Andersen. 2003. Measurement of quality of life VI. Quality-adjusted life years (QALY) is an unfortunate use of the quality-of-life concept. *TheScientificWorldJournal* 3 (816208):1015–9. doi: [10.1100/tsw.2003.79](https://doi.org/10.1100/tsw.2003.79).
- Verma, S., and J. Rubin. 2018. Fairness definitions explained. In *FairWare '18: Proceedings of the international workshop on software fairness*, 1–7. New York: Association for Computing Machinery.
- Vollmer, S., B. A. Mateen, G. Bohner, F. J. Király, R. Ghani, P. Jonsson, S. Cumbers, A. Jonas, K. S. L. McAllister, P. Myles, et al. 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 368 (l6927):1–12. doi: [10.1136/bmj.l6927](https://doi.org/10.1136/bmj.l6927).
- Walsh, E. 2020. Cognitive transformation, dementia, and the moral weight of advance directives. *The American Journal of Bioethics* 20 (8):54–64. doi: [10.1080/15265161.2020.1781955](https://doi.org/10.1080/15265161.2020.1781955).
- Wang, M., C. Xia, L. Huang, S. Xu, C. Qin, J. Liu, Y. Cao, P. Yu, T. Zhu, H. Zhu, et al. 2020. Deep learning-based triage and analysis of lesion burden for COVID-19: A retrospective study with external validation. *The Lancet Digital Health* 2 (10):e506–e515. doi: [10.1016/S2589-7500\(20\)30199-0](https://doi.org/10.1016/S2589-7500(20)30199-0).
- Wasserman, J. A., and M. C. Navin. 2018. Capacity for preferences: Respecting patients with compromised decision-making. *The Hastings Center Report* 48 (3):31–9. doi: [10.1002/hast.853](https://doi.org/10.1002/hast.853).
- Watson, D. S., J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes, and L. Floridi. 2019. Clinical applications of machine learning algorithms: Beyond the black box. *BMJ* 364 (1886):1–4. doi: [10.1136/bmj.l886](https://doi.org/10.1136/bmj.l886).
- Wilm, S., V. Leve, and S. Santos. 2014. Is it quality of life that patients really want? Assessment from a general practitioner's perspective. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 108 (2–3): 126–9. doi: [10.1016/j.zefq.2014.03.003](https://doi.org/10.1016/j.zefq.2014.03.003).
- Wolf, S. M., ed. 1996. *Feminism & bioethics: Beyond reproduction*. New York: Oxford University Press.
- Woopen, C. 2014. The significance of quality of life—an ethical approach. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 108 (2–3):140–5. doi: [10.1016/j.zefq.2014.03.002](https://doi.org/10.1016/j.zefq.2014.03.002).
- World Health Organization. 2012. *WHOQOL user manual*. Geneva: World Health Organization.
- World Health Organization. 2021. *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization.