Topoi (2023) 42:857–866
https://doi.org/10.1007/s11245-023-09911-8

Ethics of AI and Health Care: Towards a Substantive Human Rights Framework

S. Matthew Liao[1]

Accepted: 17 March 2023 / Published online: 12 April 2023

Abstract

There is enormous interest in using artificial intelligence (AI) in health care contexts. But before AI can be used in such settings, we need to make sure that AI researchers and organizations follow appropriate ethical frameworks and guidelines when developing these technologies. In recent years, a great number of ethical frameworks for AI have been proposed. However, these frameworks have tended to be abstract and not explain what grounds and justifies their recommendations and how one should use these recommendations in practice. In this paper, I propose an AI ethics framework that is grounded in substantive, human rights theory and one that can help us address these questions.

Keywords Artificial intelligence · Ethics · Health care · Human rights · Fundamental Conditions Approach

1 Introduction

There is enormous interest in using artificial intelligence (AI) in healthcare contexts. AI applications have begun to diagnose some types of cancer better than doctors,[1] identify heart rhythm abnormalities like cardiologists,[2] diagnose various eye diseases as well as ophthalmologists,[3] and identify viable embryos as fertility specialists do.[4] But before AI is used in healthcare settings, we should make sure that companies and AI researchers follow appropriate ethical frameworks and guidelines when developing these technologies. In recent years, a great number of ethical frameworks for AI have been proposed. To date, there are over 80 such frameworks from private companies, governmental agencies, academic and research institutions, and intergovernmental and other organizations.[5] These frameworks have some recommendations in common. For instance, many draw on the four principles of biomedical ethics: autonomy, beneficence, non-maleficence, and justice.[6] Among other things, autonomy seeks to ensure that patients and consumers are fully informed of, and understand, the risks and benefits of a particular health AI technology, and voluntarily consent to it. Beneficence aims to guarantee that AI health applications promote the well-being of patients and that of society as a whole. Non-maleficence strives to ensure that health AI technologies do not impose undue harm on patients. Justice seeks to promote the fair and equitable distribution of the benefits and burdens of AI health technologies among

[1] David Capper et al., "DNA Methylation-Based Classification of Central Nervous System Tumours," *Nature* 555, no. 7697 (2018).

[2] A. Y. Hannun et al., "Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network," *Nat Med* 25, no. 1 (2019).

[3] Wei Lu et al., "Applications of Artificial Intelligence in Ophthalmology: General Overview," *Journal of Ophthalmology* 2018 (2018).

[4] Renjie Wang et al., "Artificial Intelligence in Reproductive Medicine," *Reproduction* 158, no. 4 (2019). AI robots are also being created to take care of embryos in artificial wombs (https://www.independent.co.uk/life-style/gadgets-and-tech/robot-nanny-china-population-b2004342.html).

[5] https://inventory.algorithmwatch.org/.

[6] Brent Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence* 1, no. 11 (2019); Jess Whittlestone et al., *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research* (London: Nuffield Foundation, 2019); Luciano Floridi et al., "AI4people-an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and machines* 28, no. 4 (2018).

S. Matthew Liao
matthew.liao@nyu.edu

[1] New York University, 708 Broadway, Floor 6, New York, NY 10003, USA

individuals and society. In addition to these four principles, many frameworks also list such recommendations as transparency, explainability, and trust, given that some forms of AI are not understood easily, if at all, even by those who program them.[7] At the same time, many organizations also offer their own distinct recommendations. For instance, the Future of Life Institute lists "value alignment," according to which "highly autonomous AI systems should be designed so that their goals and behaviors are aligned with human values," as one of its recommendations.[8] Or, Microsoft recommends "inclusiveness," according to which "AI systems should empower everyone and engage people."[9]

In one sense, it is welcome news that industry, state, and academic institutions were concerned enough with the ethical design and use of AI to put forward these frameworks and guidelines. In another sense, however, this proliferation of ethical frameworks has created confusion, from which pressing questions arise. How did these industry, states, and academic institutions arrive at these particular sets of recommendations and not others? Which recommendations should AI developers and organizations follow and why? More fundamentally, what grounds and justifies these recommendations? How do we distinguish between recommendations that are genuine ethical principles from those that are not? Furthermore, how does one use these recommendations in practice? For instance, it seems reasonable that we should not impose undue harm on patients. But how should we achieve this? Likewise, it seems reasonable that we should be able to trust an AI system. But how do we decide which AI systems to trust?

Unfortunately, most of these AI ethics frameworks have been silent on these questions. As a result, they have been criticized for offering abstract, high-level principles that in practice have provided little concrete guidance.[10] Moreover, some have expressed concern that these frameworks are merely forms of "ethics washing"[11] and virtue-signaling,[12] where organizations and companies exaggerate their interest in ethical AI as a public relations exercise and perhaps to forestall governmental regulation. To stave off such accusations, we therefore need an AI ethics framework that is grounded in substantive normative theory, one that can help us assess whether a recommendation is a genuine ethical principle or not, and one that can give us more concrete guidance.

Elsewhere, I have developed what I call the Fundamental Conditions Approach to human rights, according to which human beings have human rights to the fundamental conditions for pursuing a good life.[13] In this paper, I shall argue that the Fundamental Conditions Approach can be extended to the use of AI in health care. I shall illustrate how this approach can help us evaluate the merits of a given recommendation. I shall also show how it can help AI researchers in health care identify distinct ethical considerations that they might encounter. To develop this framework, let me first say something about what AI is and how current forms of AI can give rise to ethical problems.

2 Machine Learning: Key Concepts and Current Limitations

The term 'artificial intelligence' is used to mean different things in different contexts.[14] For our purpose, we can understand AI broadly as getting machines to do things that, when intelligent beings such as humans do them, require cognitive functions such as thinking, learning, and problem solving. On this understanding of AI, there are different subtypes of

[7] "AI4people-an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.".

[8] https://futureoflife.org/ai-principles/.

[9] https://www.microsoft.com/en-us/ai/responsible-ai.

[10] Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines* 30, no. 1 (2020); Luciano Floridi, "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical," *Philosophy & Technology* 32, no. 2 (2019).

[11] "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical.".

[12] Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI.".

[13] S. Matthew Liao, "Human Rights as Fundamental Conditions for a Good Life," in *Philosophical Foundations of Human Rights*, ed. Rowan Cruft, S. Matthew Liao, and Massimo Renzo (Oxford: Oxford University Press, 2015); S. Matthew Liao and Collin O'Neil, "The Grounds of Ancillary Care Duties," in *Current Controversies in Bioethics*, ed. S. Matthew Liao and Collin O'Neil (New York: Routledge, 2017).

[14] This section draws on S. Matthew Liao, "A Short Introduction to the Ethics of Artificial Intelligence," in *Ethics of Artificial Intelligence*, ed. S. Matthew Liao (New York: Oxford University Press, 2020).

AI.[15] One type is Symbolic AI, or Good-Old-Fashioned Artificial Intelligence (GOFAI), which uses a series of explicitly programmed if–then rules and statements to establish the relations between inputs and outputs.[16] Another type of AI is machine learning, which uses algorithms to learn from data without being explicitly programmed to do so. Within machine learning, one can distinguish between supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, an algorithm is trained on a data set in which the correct answers for certain data are known and the data are labeled accordingly. Once the algorithm learns the relationship between inputs and outputs, it can then apply what it has learned to predict the correct answer in different (target) data sets. In unsupervised learning, a given data set has not been labeled, and an algorithm aims to sort the data on its own. In reinforcement learning, a reinforcement learning agent attempts to learn through experience.[17] Reinforcement learning algorithms work by rewarding an agent if it succeeds in a task and/or punishes the agent if it fails. Through trial-and-error, the agent strives to maximize the long-term reward. These methods also can be combined with deep learning, which uses different layers of nodes to detect increasingly abstract features, which maximize information capture while minimizing losses in predictive accuracy.

As impressive as machine learning is, it also suffers certain limitations. While many of these limitations are not unique to machine learning, they can give rise to a host of ethical issues, which are important to keep in mind when developing a substantive ethical framework for AI. First, machine learning needs a lot of data to work well. For example, supervised learning algorithms can fine-tune themselves and achieve great predictive power when they have access to a vast amount of data. Consequently, this incentivizes companies and organizations to harvest or buy data, including sensitive personal data, even when doing so might violate an individual's right to privacy. One example of this is when the Royal Free NHS Foundation Trust provided the personal data of about 1.6 million patients to Google DeepMind in 2017 to test a novel way of detecting kidney injuries without properly informing the patients about how their health data will be used.[18] Another example of this might be when the drug maker GlaxoSmithKline bought the exclusive rights to mine the genetic data of customers of the DNA testing service 23andMe for drug discovery.[19]

Second, machine learning is only as good as the data from which it learns. If a machine learning algorithm is trained on inadequate or inaccurate data, then the algorithm will make bad predictions even if it itself is well designed. For instance, algorithms trained on gender-imbalanced medical imaging datasets have been found to do worse at reading chest x-rays for an underrepresented gender.[20] Similarly, there are reasons to be concerned that skin-cancer detection algorithms may not do as well detecting skin cancer affecting people with darker skin, because many of these algorithms are trained primarily on light-skinned individuals.[21]

Third, even if a machine learning algorithm receives adequate and accurate data, if the algorithm itself is bad, it will also make bad predictions. For instance, a bad machine learning algorithm may identify a pattern even if there isn't one, a problem known as "overfitting,"[22] or may fail to identify a pattern even when there is one, a problem known as "underfitting."[23] A machine learning algorithm may also give too much or too little weight to certain features or fail to include certain relevant features altogether. Faulty algorithms can have serious ethical implications. For example, an algorithm used widely in US hospitals to determine which patients should get extra care was found to discriminate against black people because it used health costs as a proxy for health needs and, owing to structural inequalities, black patients often spend less on health care than white patients. As a result, the algorithm falsely concluded that black patients were healthier than equally sick white patients.[24] Or, in 2016 the Arkansas Department of Human Services began to use an algorithmic tool developed by interRAI to determine how many hours of home care some people with disabilities should receive.[25] The

[15] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Prentice Hall Press, 2010), which provides a good overview of different types of AI algorithms. For another perspective, see Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (New Yrok: Basic Books, 2015).

[16] John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, Massachusetts: MIT Press, 1985).

[17] Russell and Norvig, *Artificial Intelligence: A Modern Approach*., Chapter 21.

[18] https://www.bbc.com/news/technology-40483202.

[19] https://www.wired.com/story/23andme-glaxosmithkline-pharma-deal/.

[20] Agostina J. Larrazabal et al., "Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-Aided Diagnosis," *Proceedings of the National Academy of Sciences* 117, no. 23 (2020).

[21] Adewole S. Adamson and Avery Smith, "Machine Learning and Health Care Disparities in Dermatology," *JAMA Dermatology* 154, no. 11 (2018).

[22] https://www.d2l.ai/chapter_multilayer-perceptrons/underfit-overfit.html.

[23] https://www.d2l.ai/chapter_multilayer-perceptrons/underfit-overfit.html.

[24] Ziad Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (2019).

[25] https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

department implemented the algorithm's recommendation to reduce drastically the number of home care hours for many beneficiaries, which caused several people to be hospitalized, because the algorithm had incorrectly coded conditions such as cerebral palsy and had not accounted for conditions such as diabetes. Ultimately, a judge ruled that the department had improperly implemented the interRAI algorithm and ordered that its use be terminated.

Fourth, as noted earlier, deep learning is a black box that raises issues such as interpretability, explainability, and trust.[26] Deep learning is impenetrable even to its programmers because it typically employs thousands or millions of connections that interact with one another in complex ways. Given this, it is difficult to interpret what those interactions mean. The issue of explainability arises because humans often need to know how a decision is reached. However, deep learning announces its prediction without explaining (in human terms) how it arrived at that prediction. To see why this could be a problem, consider the following example. Suppose that a deep learning algorithm predicts that there is a 70% chance that Jay's tumor will become malignant in 5 years. The deep learning algorithm does not, for example, say, "There is a 70% chance that Jay's tumor will become malignant in 5 years because Jay has a history of cancer, is over 50 years old, and has lower back pain."[27] Without such an explanation, the doctor would be unable to explain to Jay why his tumor is likely to become malignant. Beyond explainability, this also raises the issue of trust in the deep learning system since we do not know whether it makes its predictions on reasonable and reliable grounds. For high-stakes decisions in health care, not being able to trust the deep learning system is especially problematic.

It might be thought that the importance of interpretability and explainability is overstated.[28] According to this line of thought, there is a trade-off between accuracy and explainability in deep learning. Given this, if a deep learning system can make accurate predictions, then it may not matter if the deep learning system is not interpretable and explainable. To support this point, one might point out that it is common for clinicians to prescribe medications such as aspirin as an analgesic and lithium as a mood stabilizer without fully understanding why they work.[29]

However, while we do not fully understand how medications work in many cases, arguably, we do have some ideas regarding the causal mechanisms through which they work. For instance, people knew that something from a willow causes fevers and pain to be reduced, even if they did not know about salicylic acid, an active ingredient in the production of aspirin.[30] This contrasts with a deep learning system which works through associations and is, at least for now, unable to track causal relations.

But it might be asked, why does it matter whether a deep learning system can track causal relations or not? One reason is that deep learning is vulnerable to certain kinds of adversarial attacks, which are inputs to machine learning models that are designed to cause the model to make a mistake."[31] For instance, deep neural networks are vulnerable to the so-called one-pixel attacks.[32] In one study, by changing just one pixel in an image, researchers were able to get a deep learning algorithm to classify an image of a car as a dog.[33] Recently, researchers have found that adversarial attacks can also be done on medical machine learning.[34] The fact that deep learning networks are vulnerable to these types of attacks suggests that deep learning networks are not learning "real" features of the world such as causal relations or what a macro-level object like a car really is; instead, these deep learning networks are only learning superficial features. For our purpose, if a deep learning network can be tricked in these ways, issues of explainability and trust remain highly relevant, especially in high-stakes domains such as medicine where human beings could be harmed.

In sum, given all the ways that machine learning could fail, it is critical that we have an ethical framework that can provide adequate guidance in these scenarios.

3 The Fundamental Conditions Approach to Human Rights

Next, let me explicate the Fundamental Conditions Approach to human rights by explaining why human beings have human rights to the fundamental conditions for pursuing a good life.

[26] Zachary C. Lipton, "The Mythos of Model Interpretability," *Queue* 16, no. 3 (2018).

[27] R. A. Deyo and A. K. Diehl, "Cancer as a Cause of Back Pain: Frequency, Clinical Presentation, and Diagnostic Strategies," *J Gen Intern Med* 3, no. 3 (1988).

[28] Lipton, "The Mythos of Model Interpretability."; Alex John London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability," *Hastings Center Report* 49, no. 1 (2019).

[29] See, e.g., "Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability.", who expresses this concern.

[30] Mohd Shara and Sidney J. Stohs, "Efficacy and Safety of White Willow Bark (Salix Alba) Extracts," *Phytotherapy Research* 29, no. 8 (2015).

[31] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv e-prints* (2014), https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6572G; N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access* 6 (2018).

[32] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi, "One Pixel Attack for Fooling Deep Neural Networks," *arXiv e-prints* (31), https://ui.adsabs.harvard.edu/abs/2017arXiv171008864S.

[33] Ibid.

[34] Samuel G. Finlayson et al., "Adversarial Attacks on Medical Machine Learning," *Science* 363, no. 6433 (2019).

As I see it, a good life is one spent pursuing basic activities such deep personal relationships with one's partner, friends, parents, children; knowledge of the workings of the world, of oneself, of others; active pleasures such as creative work and play; and passive pleasures such as appreciating beauty. The fundamental conditions for pursuing a good life are various goods, capacities, and options that human beings *qua human beings* need to pursue these basic activities.[35] For example, the fundamental goods are resources that human beings qua human beings need in order to sustain themselves corporeally, including food, water, and air. The fundamental capacities are powers and abilities that human beings qua human beings require in order to pursue the basic activities. These capacities include the capacity to think, to be motivated by facts, to know, to choose an act freely (liberty), to appreciate the worth of something, to develop interpersonal relationships, and to have control of the direction of one's life (autonomy). The fundamental options are those social forms and institutions that human beings qua human beings require if they are to be able to exercise their essential capacities to engage in the basic activities.

The fundamental conditions for pursuing a good life ground human rights because having these conditions is of fundamental importance to human beings, and because rights can offer powerful protection to those who possess them. The former is true because if anything is of fundamental importance to human beings, then pursuing a characteristically good human life is. It seems clear that if we attach a certain importance to an end, we must attach this importance to the (essential) means to this end. Given this, since pursuing a good life is of fundamental importance to human beings, having the fundamental conditions for pursuing a good life must also be of fundamental importance to human beings.

That rights can offer powerful protection to those who possess them is well known.[36] By their nature, rights secure the interests of the rightholders by requiring others, the duty-bearers, to perform certain services for the rightholders or not to interfere with the rightholders' pursuit of their essential interests. In addition, at least on certain structural accounts of rights, rights typically prevent the rightholders' interests that ground rights from being part of a first-order utilitarian calculus. This means that if a rightholder has a right to something, V, then typically no non-right claims can override the rightholder's right to V.[37] Finally, as some writers have pointed out, because the rightholders are entitled to these services as a matter of rights, this means that the rightholders can simply expect the services without requesting them.[38] Given the strong protection that rights can offer for the rightholders, and given the importance of having these fundamental conditions to human beings, it seems reasonable that human beings have rights to these fundamental conditions. If this is correct, this explains why human beings have human rights to the fundamental conditions for pursuing a good life.

The Fundamental Conditions Approach can explain why many of the recommendations found in various AI ethics frameworks are genuine ethical principles. For instance, consider autonomy, found in many such frameworks, which requires that patients be fully informed of, and understand, the risks and benefits of a particular AI health application, and that they voluntarily consent to it. The Fundamental Conditions Approach can readily explain and justify this principle. As noted earlier, autonomy, understood as being able to control the direction of one's life, is one of the fundamental conditions. To be able to control the direction of one's life in the context of AI healthcare, one needs to be informed of, and understand, the risks and benefits of a

[35] My notion of fundamental conditions might prompt some to think of Martha Nussbaum's Central Capabilities Approach Martha C. Nussbaum, *Creating Capabilities: The Human Development Approach* (Cambridge: Belknap Press, 2011). In Liao, "Human Rights as Fundamental Conditions for a Good Life.", I explain in greater detail how the two views differ. All too briefly, the hallmark of Nussbaum's approach is her emphasis on our opportunities to choose to do certain things, i.e., capabilities, rather than on what we actually choose to do, i.e., functionings. However, many human rights cannot be adequately explained in terms of capabilities. For example, in the Universal Declaration of Human Rights, there are a number of human rights that protect our moral status as persons, i.e., status rights, such as the right to recognition everywhere as a person before the law (Article 6); the right to equal protection before the law (Article 7); the right against arbitrary arrest, detention or exile (Article 9); the right to a fair and public hearing (Article 10); the right to be presumed innocent until proven guilty (Article 11). Nussbaum's approach seems to imply that one can sometimes choose not to exercise these rights, since capabilities are concerned with our real opportunities to choose. But it does not seem that one can sometimes choose whether or not to exercise these rights. For instance, it does not seem that one can sometimes choose not to be recognized everywhere as a person before the law; choose not to have equal protection before the law; choose to be arrested arbitrarily; choose to have an unfair hearing; and choose to be presumed guilty. Hence, capabilities do not seem particularly well-suited to explain these rights. In contrast, my approach can explain status rights. When we pursue the basic activities, conflicts with others are bound to arise. If and when such conflicts arise, we need guarantees that we would be treated fairly and equally. Fair trial, presumption of innocence, equal protection before the law, not arrested arbitrarily, and so on serve to ensure that we would be treated fairly and equally. As such, they are things that human beings qua human beings need whatever they qua individuals might need in order to pursue the basic activities. As such, the approach I advocate can explain why there are these human rights.

[36] Rights could also have non-instrumental importance in addition to having instrumental importance.

[37] Ronald Dworkin, *Taking Rights Seriously* (London: Duckworth, 1977).

[38] Joel Feinberg, "The Nature and Value of Rights," in *Bioethics and Human Rights: A Reader for Health Professionals*, ed. Elsie L. Bandman and Bertram Bandman (Boston: Little, Brown, 1970).

particular AI health application, and one's use of this technology needs to be voluntary. The Fundamental Conditions Approach therefore implies that patients have a right to have sufficient information and the time to decide whether to use a particular AI health application, and a right to make that decision without being coerced or exploited.

Likewise, consider non-maleficence, which is also found in many frameworks, and seeks to ensure that AI health applications do not impose undue harm on patients and that risks of harm are minimized and can be justified. Again, the Fundamental Conditions Approach can readily explain and justify such a principle. If patients were to experience harm when using a particular AI health application, this would undermine their ability to pursue the basic activities. The Fundamental Conditions Approach therefore implies that patients have a right not to have such harm imposed on them unnecessarily, which means that companies and AI researchers should do whatever they can to minimize the risks of harm to patients.

At the same time, the Fundamental Conditions Approach would also exclude some of the recommendations as genuine ethical principles. To give an example, consider "value alignment," according to which "highly autonomous AI systems should be designed so that their goals and behaviors are aligned with human values."[39] According to advocates of value alignment, what this means is that AI systems should be designed so that they can learn the correct human values from observing examples of human behavior.[40] Many people endorse this recommendation because they are concerned that AI systems will soon outpace humans and they want to ensure that algorithms are designed in such a way that will not harm to humanity. However, it is not clear that AI systems should learn from observing examples of human behavior, given that human values vary widely, and only some of them are good. Indeed, while many people would regard Mother Teresa as a moral exemplar, there are others who would not and who would instead regard racists as moral exemplars. Nevertheless, AI systems should not be designed so that their goals and behaviors are aligned with the values of those who prefer racists.

A more plausible approach in the vicinity may be to design AI systems in such a way that their goals and behaviors *respect persons or humanity as ends in themselves*.[41] The Fundamental Conditions Approach can explain why this would be a genuine ethical principle, since having our moral status as persons respected is a fundamental condition for pursuing the basic activities. Indeed, if our moral status as persons were not respected, then others would be at liberty to use us a mere means to their own end. If so, we would not have the kind of control necessary to determine the direction of our lives. In any case, the Fundamental Condition Approach implies that the goal should not be ensure that AI systems align with human values but instead to make sure that they can and do respect the value of humanity.

4 Applying the Fundamental Conditions Approach to Healthcare Algorithms

The Fundamental Conditions Approach does not just give us a substantive ethical framework for determining which recommendations are genuine ethical principles and for explaining why they are genuine ethical principles. As I shall now illustrate, it also offers AI researchers and organizations in health care a helpful framework for identifying distinct ethical considerations that they might encounter and for explaining and justifying those ethical considerations.

To illustrate, in health care, there is a spectrum in which healthcare algorithms can be deployed, ranging from inside the human organism to outside of it. The former might include a smart pill injected into the body to monitor vital signs, while the latter might involve processing data from an imaging device that detects skin cancer. The fact that a particular healthcare algorithm will be placed inside a human organism can raise distinct ethical considerations.

Here is one consideration: such an algorithm can directly and negatively impact a human being's basic health. Basic health is the adequate functioning of the various parts of our organism that are needed for the development and exercise of the fundamental capacities such as the capacity to think.[42] For instance, various life processes (including respiration, digestion, absorption, metabolism, circulation) and organ systems (including the nervous system, the skeletal system, the cardiovascular system, the digestive system, the immune system, and the reproductive system) make up, enable, and sustain these fundamental capacities. These parts of our organism must function adequately for us to develop and exercise the fundamental capacities.

It should be clear that algorithms that operate inside the human organism can directly and negatively impact our important life processes. Indeed, a smart pill inside one's body that uses AI algorithms to determine what kind of drugs to administer and when could deliver the wrong drugs

[39] https://futureoflife.org/ai-principles/.

[40] See, e.g., Stuart Russell, "Artificial Intelligence: A Binary Approach," in *Ethics of Artificial Intelligence*, ed. S. Matthew Liao (New York: Oxford University Press, 2020).

[41] For some issues with trying to encode human values that are broadly thought to be ethical into an AI system, see, e.g., Isaac Asimov, "Runaround," (3); Liao, "A Short Introduction to the Ethics of Artificial Intelligence.".

[42] See also "Health (Care) and Human Rights: A Fundamental Conditions Approach," *Theoretical Medicine and Bioethics* 37, no. 4 (2016).

or deliver the right drugs but at the wrong time, thereby disrupting and possibly damaging various life processes.

For our purpose, the Fundamental Conditions Approach can explain why basic health is an ethical consideration that we should take seriously. Basic health is something that human beings qua human beings need whatever else they qua individuals might need in order to pursue a good life. Indeed, without basic health, human beings would not possess the fundamental capacities; and without possessing the fundamental capacities, human beings would be unable to pursue a good life. As such, basic health is a fundamental condition for pursuing a good life. Earlier, we said that human beings have human rights to the fundamental conditions for pursuing a good life. It follows that human beings have a human right to basic health. Hence, the Fundamental Conditions Approach tells us to take basic health seriously because it is a human right. That human beings have this right means that they have a right not to be exposed to an undue risk of something such as a healthcare algorithm that can negatively impacting their basic health.

Here's another consideration. To get healthcare algorithms to operate inside a human organism, we would need to put them inside someone's body. However, bodily integrity is also a fundamental condition for pursuing a good life because without control over one's body, a human being would not be able to pursue a good life. Hence, on the Fundamental Conditions Approach, we also have a human right to bodily integrity. The Fundamental Conditions Approach offers another reason why we should be more cautious when deploying healthcare algorithms inside a human body, namely, because there is the potential to undermine someone's right to bodily integrity.

To give another example of how the Fundamental Conditions Approach can help identify distinct ethical considerations that AI researchers in health care might encounter, there is another spectrum in which healthcare algorithms can be deployed, ranging from those designed to assist human beings with their decision-making to those capable of making decisions on their own. An example of the former might be algorithms that can recommend potential medical diagnoses. Physicians can then elect to incorporate these algorithmic diagnoses into their decisions about how to treat patients. An example of the latter might be a robot surgeon designed to make incisions and perform surgery without any human input or control.

With respect to healthcare algorithms that enhance but do not replace human decision-making, we remain in control of how they are used. This means that we can choose not to use these algorithms if doing so happens not to align with our preferences or if we believe that doing so could cause harm and violate the rights of others. However, with respect to algorithms that make their own decisions independently, we may no longer be fully in control of them once they are deployed. Among other things, this can result in these algorithms inadvertently going against our preferences and exposing others to harm without our being able to stop them. The Fundamental Conditions Approach implies that algorithms that can operate autonomously would require distinct scrutiny since they could inadvertently subject others to harm and violate the rights of others without our inputs.

Interestingly, healthcare algorithms could be deployed on both spectrums at same time, resulting in at least four types of "combined" algorithms:

Type I Algorithms that operate inside a human organism and make their own decisions;

Type II Algorithms that operate outside a human organism and make their own decisions;

Type III Algorithms that operate inside a human organism and serve as inputs in human decisions; and

Type IV Algorithms that operate outside a human organism and serve as inputs in human decisions.

We can place these four types of algorithms in a 2 × 2 matrix.

	Serve as inputs in human decisions	Make their own decision
Inside human organism	Type III	Type I
Outside human organism	Type IV	Type II

Some examples of Type I might include next generation Brain-Computer Interfaces (BCIs) or smart pills, both of which would operate inside a human organism and be able to make independent decisions without human input. Some examples of Type II might include next generation robot surgeons or robot caretakers, both of which would operate outside a human organism but could make decisions on their own without human input. An example of Type III might be AI-powered in-vivo biosensors that can continuously monitor biological processes inside a human organism and provide this information to physicians for further analysis. An example of Type IV might be AI-enabled radiology assistants, which can improve diagnostic accuracy.

For our purpose, the Fundamental Conditions Approach can help us see that the ethical considerations we have identified earlier could remain even when the algorithms are deployed on both spectrums. To give an example, consider a next generation BCI that would operate inside a human organism and be able to make independent decisions without human input. The Fundamental Conditions Approach tells us that we should make sure that such a device would not violate the right to basic health and the right to bodily integrity, and also that it should not inadvertently cause harm to others.

I will now show that being able to identify the kinds of ethical considerations that one might encounter with respect to a particular healthcare algorithm can help us decide how it should be developed and implemented.

5 A Case for Locking Some Healthcare Algorithms

As we have seen from the section on the current limitations of machine learning, a pressing problem with current iterations of deep learning systems is that their learning is in some sense ‘superficial,’ that is, they cannot grasp causal relations and may not learn about real features of the world. As such, deep learning is prone to getting things seriously wrong, as its vulnerability to adversarial attacks suggests. What can we do to reduce the risk of algorithms going astray in the health care context? There are at least two options.

The first option is to hold the algorithms fixed so that they would give the same results whenever they are provided with the same inputs or, as the FDA puts it, use “locked algorithms.”[43] By way of contrast, “adaptive algorithms” are able to learn continuously, which means that for a given set of inputs, the outputs may change as the learning process is updated.

A second option is to hold the environment in which the algorithms operate fixed and allow for the use of adaptive algorithms. For instance, consider next-generation robotic surgeons. Suppose that we would like to use adaptive algorithms in such robotic surgeons. We might be able to reduce the risks of the algorithms’ going astray by holding the environment in which they operate fixed. For instance, we might impose fixed parameters and allow a robotic surgeon only to perform tasks that it can perform to a high degree of accuracy such as incisions or suturing.

However, in many healthcare applications, it may be difficult to hold fixed the environment in which algorithms operate, since many such applications involve the human organism, the life processes of which are in constant flux and therefore difficult to hold fixed. Given this, it seems that the first option—locking the algorithm itself—may be preferable for many types of healthcare applications at least in the near term. For instance, earlier we identified four types of algorithms, where Type I involves algorithms that operate inside a human organism and make their own decisions; Type II involves algorithms that operate outside a human organism and make their own decisions; Type III involves algorithms that operate inside a human organism and serve as inputs in human decisions; and Type IV involves algorithms that operate outside a human organism and serve as inputs in human decisions. Based on the ethical considerations we have identified using the Fundamental Conditions Approach, it seems that there are some good reasons to use locked algorithms at least with respect to Types I, II, and III.[44] For example, suppose that a healthcare algorithm would be operating inside a human organism and be making its own decisions. To ensure that the basic health and bodily integrity of patients remains uncompromised and that they are not inadvertently harmed, it seems that, other things being equal, there is a good reason to use algorithms that provide the same output whenever they are given the same input, i.e., locked algorithms. Likewise, there is a *prima facie* reason to use locked algorithms in cases where algorithms outside of the human organism would be making their own decisions, since they could go against our preferences or cause harm without our being able to stop them. Similarly, in cases in which the algorithms would simply assist human decision-making but would do so inside the human organism, there is also a *prima facie* reason to use locked algorithms, since they could affect the patients’ basic health and bodily integrity.

However, even though there is a good reason to use locked algorithms in these cases, we may still be able to take advantage of adaptive algorithms through what might be called staggered learning. Staggered learning involves allowing adaptive algorithms to learn and generate new input–output relations but not apply that new learning synchronously. Once the new connections between inputs and outputs have been verified and validated, they could be used to develop a new, updated, locked algorithm. In this way, learning could still occur, but would be done in steps.

6 Conclusion

Countless private companies, governmental agencies, academic institutions have proposed ethical frameworks for AI, but they neither explain how recommendations in their frameworks are justified nor the means by which we might distinguish genuine ethical principles and those that are not genuine ethical principles. In this paper, I argued that the Fundamental Conditions Approach to human rights gives us a more unified and substantive AI ethics framework that can help us address these issues. In addition, I proposed that the Fundamental Conditions Approach offers a helpful framework for identifying distinct ethical considerations that AI

[43] U.S. Food and Drug Administration, “Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (Samd). Discussion Paper and Request for Feedback.” (2019).

[44] As far as I can tell, the ethical considerations we have identified earlier do not seem to apply to Type IV algorithms, which serve as mere inputs into human decision-making and operate outside of the human organism. This of course does not mean that there are not other ethical considerations that apply to Type IV algorithms.

researchers in health care might encounter and for explaining and justifying those ethical considerations. For instance, I showed that the Fundamental Conditions Approach helps us to see that healthcare algorithms that operate inside the human organism could raise issues about basic health and bodily integrity, and healthcare algorithms that make decisions on their own without human input could raise concerns that such algorithms could harm others without our being able to stop them. Since current iterations of deep learning learn superficially, I also proposed that the Fundamental Conditions Approach implies that many healthcare algorithms should be locked at least for now, but that we can still take advantage of adaptive algorithms through staggered learning. The Fundamental Conditions Approach offers a novel, substantive ethical framework for research in AI and health care, and deserves to be investigated further in future discussions on this topic.[45]

Funding The author did not receive support from any organization for the submitted work.

Declarations

Conflict of interest The author declares that he has no competing interests.

References

Adamson AS, Smith A (2018) Machine learning and health care disparities in dermatology. JAMA Dermatol 154(11):1247–1248

Akhtar N, Mian A (2018) Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 6:14410–14430

Asimov I (1942) Runaround

Capper D, Jones DT, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C et al (2018) DNA methylation-based classification of central nervous system tumours. Nature 555(7697):469–74

Deyo RA, Diehl AK (1988) Cancer as a cause of back pain: frequency, clinical presentation, and diagnostic strategies. J Gen Intern Med 3(3):230–8 (**In eng**)

Domingos P (2015) The master algorithm: how the quest for the ultimate learning machine will remake our world. Basic Books, New York

Dworkin R (1977) Taking rights seriously. Duckworth, London

Feinberg J (1970) The nature and value of rights. In: Bandman EL, Bandman B (eds) Bioethics and human rights : a reader for health professionals. Little, Brown, Boston, pp 19–31

Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS (2019) Adversarial attacks on medical machine learning. Science 363(6433):1287–1289

Floridi L (2019) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 32(2):185–193

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C et al (2018) AI4people-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 28(4):689–707 (**In eng**)

Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv e-prints. https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6572G

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Minds Mach 30(1):99–120

Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY (2019) Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 25(1):65–69 (**In eng**)

Haugeland J (1985) Artificial intelligence: the very idea. MIT Press, Cambridge

Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E (2020) Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci 117(23):12592–12594

Liao SM (2015) Human rights as fundamental conditions for a good life. In: Cruft R, Liao SM, Renzo M (eds) Philosophical foundations of human rights. Oxford University Press, Oxford, pp 79–100

Liao SM (2016) Health (care) and human rights: a fundamental conditions approach. Theoret Med Bioethics 37(4):259–274

Liao SM (2020) A short introduction to the ethics of artificial intelligence. In: Liao SM (ed) Ethics of artificial intelligence. Oxford University Press, New York, pp 1–42

Liao SM, O'Neil C (2017) The grounds of ancillary care duties. In: Liao SM, O'Neil C (eds) Current controversies in bioethics. Routledge, New York, pp 29–42

Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):31–57

London AJ (2019) Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 49(1):15–21

Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y (2018) Applications of artificial intelligence in ophthalmology: general overview. J Ophthalmol 2018:5278196

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1(11):501–507

Nussbaum MC, Capabilities C (2011) The human development approach. Belknap Press, Cambridge

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453

Russell S (2020) Artificial intelligence: a binary approach. In: Liao SM (ed) Ethics of artificial intelligence. Oxford University Press, New York, pp 327–341

Russell S, Norvig P (2010) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall Press, Hoboken

Shara M, Stohs SJ (2015) Efficacy and safety of White Willow Bark (Salix Alba) extracts. Phytother Res 29(8):1112–1116

[45] I would like to thank Nicholas Tilmes, an anonymous reviewer at *Topoi*, and audiences at the Association for Practical & Professional Ethics, Saint Louis University, CUNY, Fudan University, University of Cincinnati, and University of Central Florida, for their very helpful comments on earlier versions of this paper.

Su J, Vargas DV, Sakurai K (2017) One pixel attack for fooling deep neural networks. arXiv e-prints. https://ui.adsabs.harvard.edu/abs/2017arXiv171008864S

U.S. Food and Drug Administration (2019) Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (Samd). Discussion Paper and Request for Feedback

Wang R, Pan W, Jin L, Li Y, Geng Y, Gao C, Chen G et al (2019) Artificial intelligence in reproductive medicine. Reproduction 158(4):139 (**In English**)

Whittlestone J, Nyrup R, Alexandrova A, Dihal K, Cave S (2019) Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Nuffield Foundation, London

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.