Towards Ethical AI for Health:
Reshaping Healthcare Wisely

Irene Dovichi

July 4, 2023

Abstract. In this essay I will present what are the main issues to be addressed in the healthcare sector to ensure an ethical integration of AI systems. The motivation behind the choice of theme is the growing experimentation and use of computer tools for health-related purposes, which I find a particularly delicate topic. I will talk about issues related to the intrinsic nature of AI technologies, accountability, and impact on people. For each subject, I will outline what are the moral values that I believe should be followed.

Keywords. Healthcare, Ethical challenges, Improvement

1 Introduction

In recent years AI has become a huge success in all sectors, offering innovative solutions, and in the medical field its potential is high as well. Currently, the most common roles for AI in medical settings are clinical decision support, imaging analysis, and optimization of health facilities. Clinical decision support tools help doctors both for the diagnosis and to make decisions about treatments, while also providing quick access to relevant patient information or research. In medical imaging, AI algorithms are useful for analyzing CT scans and X-rays, and sometimes they can recognize patterns that human radiologists miss. However, there are many controversial aspects related to the use of these technologies that have emerged very soon. In this regard, I will present a categorization of the main ethical risks of AI in healthcare, referring to the findings in [1], [2], and

[3]. There will be three sections dedicated to different subjects: implementation challenges (2), accountability (3), identity and individuality issues (4). In the first one, I will talk about the intrinsic nature of AI systems, identifying two problems: dataset limitations, and explainability. Subsequently, I will discuss the issue of attributing responsibility, reflecting on the fact that in certain situations it is unclear who to blame, and that, in general, individuals are more involved. Lastly, I will reflect on the impact on individuals and society: in addition to the issue of privacy, the relationship between healthcare professionals and patients could radically change, and finally, at a global level, there is a risk of exacerbating the differences between advanced and developing countries. I strongly believe that, in order to fully exploit the power of information technology, clear moral values must be taken into account. In this spirit I am going to outline the key values with which to operate in the various contexts. As a general principle, I think the ultimate goal of using advanced tools in the medical field should be to improve the health of human beings and ensure equality.

2 Implementation challenges

Despite the evidence that some AI techniques are successful in a wide variety of medical tasks, there are a number of issues related to the intrinsic nature of AI systems and the availability of resources. The problems of overfitting, and the lack of reproducibility and validity of results question the scientific rigor of such tools. Therefore, we have to investigate knowledge concerns, asking ourselves how to deal with evidence that is misleading, inconclusive, or inscrutable. More pragmatically, we need to recognize that the high costs of these technologies can exacerbate inequalities between advanced and developing countries, causing only a fraction of humanity to flourish. I identified two main problems related to the implementation of AI algorithms: imperfect datasets and explainability. I tried to discuss recent progresses in the field that could represent possible solutions for the presented issues, choosing the most relevant in my opinion, and highlighting their weak points.

2.1 Dataset limitations

In this section we discuss the impact of the datasets used to train AI algorithms. In the first two modules we refer to the specific case in which AI is used for the analysis of medical images, such as radiographs, histology, and optic fundi. To give you some context, we are considering image classification problems, tackled with supervised learning algorithms trained on labeled data.

Memory. The equipment needed to obtain the inputs for AI systems, like whole-slide images, can be prohibitively expensive. Furthermore, the large size of the images taken as input by the neural networks creates a problem related to the available memory. A possible solution is that an expert identifies a region of interest in the slide and proceeds to crop the image before giving it as input. Alternatively, the pictures can be compressed or split into small parts, but in the first case the image quality will deteriorate, and in the second case the system will struggle to identify correlations between the different pieces.

Labeling. Another issue worth discussing is the data labeling process in supervised learning. Normally, the training dataset is associated with labels, from which machine learning models, such as neural networks, can learn which label to attribute to a certain input. In many cases, data labeling tasks require human intervention, and in the specific case of health data, medical experts are required for the classification process. Needless to say, this procedure is time-consuming and constrictive, since there is a shortage of available experts. Alternative data labeling approaches include crowdsourcing and delegation to specialized companies. The first method is fast, but produces less accurate labels and there are privacy concerns as the data has to be shared with many labelers, while the second method ensures quality, but is expensive. As a result, well organized datasets are difficult to obtain, and alternative methodologies are gaining interest. In particular: *semi-supervised learning* (which uses a training dataset made of a small amount of labeled data, and a large amount of unlabeled data) and *unsupervised learning* (trained with unlabeled data). In addition to being faster, these methods can reveal new patterns in disease manifestation; however,

the results are generally less accurate.

Algorithmic bias. Another big challenge that AI has to address in general is that of algorithmic bias. In fact, most AI algorithms need big databases to be trained, but several groups of people (such as gender and ethnic minorities, children and the elderly, and people with disabilities) are absent or misrepresented in the existing biomedical datasets. This can lead to misdiagnoses and the amplification of inequities for these human beings. In some cases, it is impossible to quantify biases in a dataset, unless their presence has been included as a metadata. Think, for example, to socioeconomic status or sexual orientation in a biomedical dataset; there is evidence that even this apparently irrelevant data is associated with brain functions and exposure to certain diseases, therefore it is important to know if our algorithm was trained on a balanced dataset. As suggested in [4] and [5], biases in medical analysis can be mitigated by removing *protected attributes*, or by *data augmentation*, in addition to collecting more representative data (which is the most desirable solution, but also the most expensive). Protected attributes are features linked to sensitive information; often they are not available for privacy reasons, but if they are available it can be enforced that they are not used to train the algorithm, so that the model's decision is not affected. The main issue is that non-protected attributes are often correlated with protected ones, and can be used to infer sensitive data. On the other hand, generating synthetic data permits to expand the amount of available labeled data for deep learning models. There are many types of data augmentation, and I think that, when experimenting, it is worth trying if one of these methods can improve the performance of the algorithm under consideration. I believe that addressing bias in datasets is a major issue because it can be crucial to ensure inclusiveness. A total elimination of bias is practically impossible to obtain, but this should not justify either an unfair result, or the absence of rules inciting more accurate systems.

2.2 Explainability

An intrinsic problem of AI tools is their black-box nature. People may be resistant to digitized health if they do not understand the principles behind them, therefore there is need to build user trust before submitting them to changes. In this regard, I find it relevant that the World Health Organization declares that: "AI should be intelligible or understandable to developers, users and regulators"[6].

Saliency methods. A practical approach to make more transparent models is *saliency methods*. They produce heat maps that highlight the regions of the medical image that most influenced the model's prediction. A big advantage of these maps is that they are intuitive to interpret, however there are some relevant problems to consider. It has been shown that an image subject to perturbations small enough for the algorithm's prediction to remain unchanged, can instead produce very different saliency maps. Moreover, trying to understand the outcome of a saliency method could represent an additional level of obscurity, complicating things instead of helping. In conclusion, I think that before using saliency maps as an evaluation tool, it is worth doing more research and more analysis of how to interpret these methods.

Communication. I also find it useful to think on an abstract level about the explainability issue; in this regard, the analysis carried out in [7] is particularly interesting, so I proceed to outline it. We know that the trade-off between model performance and model clarity is difficult to overcome, but transparency should not be mistaken for explainability. In fact, explaining something is a matter of communication, while transparency is an intrinsic characteristic of some model, achievable with modifications in the system. In this sense, it is better to focus on making humans understand the function of AI tools, rather than questioning about the internal structure of algorithms. The ultimate goal of fair AI systems should be increasing control by enabling a communication between users and algorithms; explanations by the system may include machine-produced verbal arguments, saliency maps, examples, and so on. Therefore, I suggest that another key principle to follow is to ensure explainability, precisely in this new sense. Namely, ensure that there is greater

comprehensibility, in a way that does not imply transparency and, therefore, does not compromise the performance of the algorithms.

3 Accountability

Another aspect that certainly cannot be overlooked is that of assigning responsibility for the use of AI systems in healthcare. In fact, even if it is not possible to explain in details how an algorithm produces its results, someone has to take responsibility for decisions made on the basis of such outputs.

Responsibility gap. As already mentioned, there is a control problem associated with AI, meaning that developers and designers cannot predict the shape that AI models take, as they evolve independently. This leads to a responsibility gap as it is not clear at what point of the process the problems arose. One possibility could be to assign responsibility to developers in case of unwanted results; this can certainly incentivize them to minimize harm to patients, but it might be unethical. Indeed, before blaming the developers, one should understand whether the unfair results arise as a consequence of some hidden assumptions they made. For example, the article [8] talks about an algorithm (widely used in America) that distributed healthcare to patients and which was found to systematically discriminate against black people. In that case, some researches found out that the developers assumed that people who spent less on healthcare were healthier. This assumption seemed reasonable, but, in practice, black people experience reduced access to care, due to distrust of the healthcare and systematic racial discrimination by healthcare professionals. In conclusion, as algorithms are designed by humans, they may often reflect *human biases*. In addition, it must be considered that, in general, many people work on a project and, therefore, it is difficult to assign responsibility. This is known as *the problem of many hands*. In this regard, one can decide to follow a faultless responsibility model, in which all the agents involved are held responsible. This approach might encourage everyone to act with integrity, but it does not take into consideration the actual intentions of each agent or their ability to

control an outcome. As a consequence, the agents involved might be reluctant to assume responsibility even for people they may not know well. Lastly, I want to briefly talk about a frequent problem in the health sector, namely that of the misdiagnosis. It could happen that a patient suffers an erroneous diagnosis because of a malfunctioning wearable device, or because a healthcare professional has relied on an algorithm that made a mistake. In the case of the personal device, responsibility could be attributed to the company that commercialized it. In the other case, it can be investigated whether the error is attributable to medical negligence (which could have carried out a more in-depth analysis beyond that of the AI system) or developer negligence (who approved a system with a rooted malfunction). All these situations are difficult to deal with, and I think there is a need to create new laws that adapt to new problems. In any case, one must try to safeguard progress, and, therefore, not put too much pressure on developers who might otherwise feel too limited.

Personal commitment. Another important aspect is that of the transition from hospital to home-based care. Smart devices make healthcare more pervasive in daily life, as they constantly give advice and instructions to patients. As a result, individuals might gain responsibility, as they are expected to actively participate in interactions with AI tools. Moreover, some people may appreciate the independence that patient-based care offers, while others may find it stressful because of the commitment it requires. The risk of victim-blaming takes hold, exposing an already sensitive category to further stress. However, sometimes there can be no certainty that the poor result is due to lack of action by the user: it could be a faulty device, buggy code, or the result of biased datasets. Speaking of this issue, I think AI technologies should be optional, so that people can decide, depending on their personalities, whether to adopt them. In any case, a key principle that companies marketing devices must follow is to protect users' autonomy.

4 Identity and individuality issues

Both on an individual and on a social level, the introduction of AI technologies has a huge impact. The risks of privacy violation and identity theft are more current than ever, but new problems are also emerging in the face of an innovative healthcare system. The integration of AI systems on a large scale can have an impact on the labour market, but also on the relationship between patients and health professionals. Furthermore, low- and middle-income countries may struggle to finance technological research and buy cutting-edge tools.

4.1 Privacy and data protection

Patient data is at the heart of any type of analysis for which AI tools can be employed. It is of primary importance to treat people's sensitive information with respect and care. Also, more information campaigns should be done about what kind of data are being collected in order to gain information on health. *Public health surveillance* consists of the systematic collection and analysis of data for planning and evaluating public health. Lately, "digital traces", such as videos watched on YouTube and internet searches, are also being added to health data [6]; this raises concerns about both the accuracy of models trained with these data, and privacy, since the principle of "purpose limitation" could be violated. In the following, we will concentrate on another issue: that of data leakage, also talking about a technique that can be used to amortize the damage.

Data leak. The digitization of healthcare would lead to the creation of datasets with sensitive information about real patients. A serious threat is that this data falls into the hands of bad actors. A possible way to prevent data leakage is to decentralize data storages. In this regard, it seems relevant to me to mention the technique of *federated learning*, whose basic principle is to reduce the sharing of data. The federated learning protocol requires developers to send AI models to various institutions that have their private datasets, then each of them trains the models independently on their own data, and once the operation is complete, sends back the updated

model. A significant limitation of this method is that it requires many devices to exchange updates frequently, which can be problematic on realistic networks that are limited in computational and communication resources. Another weakness is that federated learning is not completely immune to privacy attacks; in fact, it could happen that the training data is reconstructed starting from the final model, or that the attackers target the communication media between the central server and users. However, there is the possibility to encrypt the inputs before using them, to protect the data.

4.2 Psychological and social impact

In this last section, we reflect on the effects of reshaping healthcare from the perspective of health professionals and patients. There are some deep issues that emerge, therefore it is important to raise awareness. And lastly, we mention a new opportunity to address the already existing social gap between advanced and developing countries.

Health professionals. The introduction of AI in the healthcare system is seen by many as a threat, also due to the risk of job losses. I personally think that this fear should be reconsidered, also in light of the fact that in many countries there is indeed a shortage of medical personnel. AI systems could take over repetitive tasks, reducing the workload of healthcare workers, and new professional profiles could also emerge. In [6], three new possible types of works are outlined: trainers (who evaluate and stress-test AI systems), explainers (who discuss the reliability of algorithms), and sustainers (who identify unintended consequences). Another kind of concern in this setting is the disqualification of practitioners, i.e., doctors will tend to accept the results of a method known to be accurate, and they will lose interest and ability over time. For example, at some point physicians may no longer be able to read an X-ray because AI systems do, and it can get to the point where operators and patients are unable to act if the system fails or is compromised. This highlights that full automation would be counterproductive. Therefore, I believe it is important to move towards collaborative setups between humans and AI systems, and also towards a cooperation between computer

scientists and doctors, considering that staying up to date on both fields would be difficult.

Patients. Changing the healthcare system can also have repercussions on patients. It is feared that the dehumanization in medical consultation may undermine trust in healthcare institutions. In fact, it is difficult to accept a diagnosis that is not adequately explained, while doctors can empathize and look for suitable ways to clarify. This is a delicate issue, both because most people do not have adequate knowledge of digital health to understand how it works, and because there is an effective limit to understanding AI tools due to their black-box nature. For these reasons, I think that at the moment the optimal solution is the consultation process remains a work done by people. In this way, we can take advantage of using advanced techniques, but also prevent users from trusting these tools too much by showing them what the actual limits are.

Social gap. The development of AI could exacerbate social inequalities and discrepancies between developing and advanced countries. Low- and middle-income countries are facing a shortage of specialised personnel and healthcare services, which undermines their independence. I want to mention a new phenomenon, known as *telemedicine*, that could alleviate this disparity. It consists in providing remote monitoring systems and virtual assistants to support patient care. These services have increased during the COVID-19 pandemic, but I believe they have potential in this context as well. In fact, such tools could be provided to poor countries to help them train medical personnel and provide assistance to the sick.

5 Conclusion

The list of problems presented in this essay is by no means exhaustive. Nonetheless, I hope I have given you an idea of which are the main areas to focus on when evaluating a remodeling of the health sector. I proceed to summarize the ethical considerations discussed in the course of this dissertation. Regarding implementation issues, I hope that over time the limitations of the current models, the mem-

ory constraint and the labeling procedure, but above all the algorithmic bias, can be overcome. Moreover, we must aspire to greater explainability of AI tools, improving communication between users and algorithms. For what concerns the accountability issue, I think that the responsibility gap should be addressed with new regulations in the various contexts in which it occurs, with particular attention to misdiagnoses. Moreover, in light of the fact that we are moving towards an home-based care, we should preserve people's autonomy, and avoid them from unnecessary stress by providing tools that do not blame them in case of poor health. Another key value to focus on is protecting patient identities and data. And, on a more abstract level, caring for the psychological impact on health professionals and patients. Finally, we should not forget that, unfortunately, there are countries in great difficulty, and a huge step to improve their condition would be to assist them in the health sector. AI offers an opportunity, unprecedented in history, to make the world a better place, and I really hope that ethical considerations such as those discussed will lead us in the future.

References

[1] D. D. Farhud and S. Zokaei. “Ethical Issues of Artificial Intelligence in Medicine and Healthcare”. In: *Iran J Public Health* 50.11 (2021). DOI: 10.18502/ijph.v50i11.7600.

[2] J. Morley, C. C.V. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi. “The ethics of AI in health care: A mapping review”. In: *Social Science & Medicine* 260 (2020). DOI: 10.1016/j.socscimed.2020.113172.

[3] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. “AI in health and medicine”. In: *Nature Medicine* 28 (2022), pp. 31–38. DOI: 10.1038/s41591-021-01614-0.

[4] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara. *Addressing bias in big data and AI for health care: A call for open science*. 2021. DOI: 10.1016/j.patter.2021.100347.

[5] M.A. Ricci Lara, R. Echeveste, and E. Ferrante. “Addressing fairness in artificial intelligence for medical imaging”. In: *Nat Commun* 13 (2022). DOI: 10.1038/s41467-022-32186-3.

[6] World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance*. 2021.

[7] E. Esposito. “Does Explainability Require Transparency?” In: *Sociologica* 16.3 (2022). DOI: 10.6092/issn.1971-8853/15804.

[8] H. Ledford. *Millions of black people affected by racial bias in health-care algorithms*. 2019. DOI: d41586-019-03228-6.