

Ethics of Artificial Intelligence in global health: explainability, algorithmic bias and trust

Angeliki Kerasidou

Introduction

Recent advances in computer science and data technologies are accelerating progress in artificial intelligence (AI), opening up exciting new possibilities in the field of healthcare. AI has the potential to disrupt and transform the way we deliver care globally. It is reputed to be able to improve the accuracy of diagnoses and treatments, and make the provision of services more efficient and effective. In surgery, AI systems could lead to more accurate diagnoses of health problems and help surgeons better care for their patients. In the context of delivering healthcare in low-and-middle-income-countries (LMICs), AI could assist further. Access to healthcare still remains a global problem and the use of AI tools could help address this by facilitating access to healthcare professionals and services, even specialist services, for millions of people. For example, a combination of advanced AI tools and mobile technology could assist less experienced healthcare professionals with diagnosing and treating patients on the ground. This way patients around the world could have access to highly skilled and specialised care, which would not have been available to them otherwise. Integrating AI into healthcare promises a plethora of benefits for different populations.

A number of practical and ethical issues considerations have already been identified in the literature that relate to the development and deployment of medical AI on the ground. Issues such as those of explainability and algorithmic bias might appear to be practical in nature but resolving them leads to questions of value, fairness and trust. In this paper, I seek to highlight these issues in the context of developing AI healthcare tools for use in LMICs. It is fair to acknowledge that these issues are neither new nor exclusive to medical AI. Yet, AI's ability to deliver on its promises depends on their resolution.

Explainability, algorithmic bias and trust

Recent advances in computer and data sciences have led to a period of renewed spring for AI. One of the areas that stands to benefit is healthcare. AI is promising to positively disrupt the way healthcare is accessed and delivered by increasing the efficiency of healthcare systems, improving the accuracy of diagnoses and effectiveness of treatments, augmenting the capacities of healthcare professionals or relieving them from tasks that could be better performed by rational, tireless machines.

Currently, the majority of AI tools are developed and trialled in higher income countries such as the UK and US, which have established healthcare systems and access to the technological infrastructure, large datasets and skilled workforce. Many of the AI health applications developed, however, could be relevant for populations around the world. The WHO recently noted that AI could help struggling healthcare systems better deal with the ever increasing costs of keeping populations healthy, and also help countries meet the OECD goal of universal access to healthcare.¹ AI could help optimise the delivery of vaccines and help health community health workers use their time more efficiently in trying to reach their populations. It could facilitate the remote diagnoses of skin injuries such as burns, and the prediction of perinatal complications such as birth asphyxia.² Even in an area such as surgery which require access to highly skilled professionals and specific technologies, AI could still facilitate care access for populations in LMICs. AI systems that use deep learning to augment computer vision and technical skills could lead to more accurate diagnoses of health problems such as skin cancer, help surgeons perform more precise incisions during operations and follow the progress of postoperative patients.³ Although, many countries still lack the necessary technological infrastructure to develop AI

for themselves, cloud computing and the widespread use of smartphones are nevertheless making it possible for such advancements to reach the most remote parts of the globe.⁴

A number of ethical issues have already been identified regarding the introduction of AI in healthcare. Some relate on how AI might transform and disrupt care on the ground, including its impact on the doctor-patient relationship and the values of trust and empathy⁵, the risk for a new type of paternalism as healthcare professionals are likely to defer decisions to AI tools^{6,7}, and of the dehumanisation of healthcare as more and more tasks will be outsourced to intelligent machines.⁸ Other issues relate primarily to the way the technology is developed and applied. These issues might appear as merely practical or technical problems that can be resolved by using the right techniques and technology development steps. And yet, the manifestation but also the resolution of these issues can have profound ethical components attached to them, and form the subject of this paper. Consider, for example, the problem of explainability or what is otherwise called, the 'black box' problem. This problem relates to AI tools that use deep learning and neural networks to optimise outcomes. Information in the form of data are fed to the AI (pictures of cancerous skin lesions), which the system uses to 'autonomously' learn and produces the correct outcomes (identify cancerous skin lesions when presented with new pictures). Yet the 'thinking process' by which these outcomes are produced is not obvious to those who use the AI, or even to those who develop it. For this reason, when mistakes are made, it is impossible to understand and investigate the reason for that mistake and rectify it. This can be particularly problematic in areas such as surgery where mistakes can have an immediate, and even tragic, impact on people's health or their lives. It is possible, however, to use AI tools that do not fall foul of the explainability problem, such as those based on decision trees, a type of machine learning algorithm that can be interpreted by humans. Arguments have been made about the importance of being able to understand and explain the process by which certain decisions are made in healthcare.⁹ The decision, however, to refrain from using deep learning and neural networks in medical AI might result in decreased efficiency of these tools. Decreased efficiency in healthcare means not only higher economic costs but also costs in human lives saved. Therefore a value judgment needs to be made as to whether being able to explain AI-led decisions –thus retaining the ability to trace back the sources of error and eliminate them– is more important than the potential to save more lives and improve the wellbeing of more people by using more advanced, but less explainable, AI technologies.

Another practical ethics issue, which is particularly relevant for the development of medical AI in LMICs, is the issue of algorithmic bias, errors built into AI systems based on incomplete or biased datasets that lead to unfair outcomes. For example, an AI tools used in the US to predict care needs of patients with complex health issues systematically underestimated the care needs of black patients.¹⁰ Addressing the issue of algorithmic bias can help ensure the clinical appropriateness and relevance of AI for the patients using it. Furthermore, it is important that AI tools used in healthcare do not sustain, perpetuate and exacerbate existing biases and inequalities embedded in healthcare tools, therapeutics and systems that see certain groups receiving substandard or inappropriate care.¹¹ Again, at its surface this problem appears to be a practical one. Algorithmic bias can be corrected by ensuring that AI tools are developed and trained using diverse and inclusive datasets, representative of the populations that stand to benefit from these applications. Yet, achieving this objective is complicated not only in practical but also ethical ways.

Once gaps in the datasets are identified, an organised effort will be necessary to ensure that data are collected from all relevant populations and communities, analysed and incorporated in the systems appropriately. Advances in data technologies in recent years, and the widespread use of mobile phones have facilitated the collection of data from a wide range of populations. However, the case

remains that certain populations, such as women in LMICs, are still less likely to have access to mobile devices and the internet¹², and are therefore underrepresented in training datasets. AI developers building tools for universal use need not only to be cognisant of such data gaps, but also to take active steps to redress it. This might require actively targeting certain populations (e.g. through a research programme), and also ensuring that data remain up to date as developers continue to train and refine their algorithms. Reaching these populations is only one of the hurdles that needs addressing. The next important step is convincing them to submit their data to the research institutions and global corporations developing these technologies. This last step necessarily touches on ethical issues surrounding trust: if people trust the stakeholders who seek access to their data they are more likely to agree to permitting them access to their information. Distrust however, could severely hinder this effort, and lead to structural flaws in training datasets that result from levels of population distrust in major institutions.

Trust, in the context of research is a complex and complicated relationship, which operates in multiple levels simultaneously involving different moral actors, from individuals, to groups, to private companies and public institutions. The issue of trust is neither new nor uniquely related to medical AI development. Trust in biomedical research and in the context of LMICs has been discussed in the bioethics literature.^{13,14} Factors such as whether there is a personal relationship with the researchers, or the reputation of research institutions (positive or negative) have shown to have an impact on reported trust.¹⁵ There is, however, a particular aspect of medical AI research and development that can bring a new dimension to the trust discussion. This is the participation of new type of companies in the development of AI tools for healthcare. The drive to developing data-driven healthcare tools and services, including medical AI, is attracting huge capital investment – which is increasingly dominated by large, private tech companies, such as Google and Amazon, which have access to vast amounts of data and the technological capacity to develop tools quickly. Empirical research conducted in Kenya regarding data sharing in health research has demonstrated that one of the main concerns of research stakeholders, including members of the public, is ensuring that the populations who contribute to scientific and technological advancements through their data should also benefit from these advancements.¹⁶ Similar concerns regarding justice and fairness have been voiced by populations in other parts of the world who have expressed their scepticism and often their distrust of private corporate interests that operate in the healthcare sector.¹⁷ Their main concern is that private corporations, the principal goal of which is the generation of profit, are less likely to act for the benefit of society as a whole. Furthermore, the fact that many interventions are developed by such commercial companies means that study results are less likely to be published in academic journals, and thus less likely to receive academic scrutiny through peer review. A result of global companies racing to dominate this emerging area of healthcare is that medical AI tools are often employed on the ground with neither adequate evidence regarding their effectiveness, nor regulatory and institutional safeguards.¹⁸ Moreover the profit motive means that such companies are more likely to target developing tools more relevant to higher-income markets, rather than focus on addressing issues specific to LMICs. All these factors could have an impact on the level of trust LMIC populations might have in medical AI and the companies that are developing them.

The solution to the issue of trust is not, or at least not only, to nudge companies towards becoming more trustworthy. Trustworthiness is a self-motivated and self-regulated attitude.¹⁹ As such companies can decide what trustworthiness should look like in their domain, and what is an appropriate way to demonstrate it. But they can also decide when trustworthiness stops being convenient or profitable, to the potential disbenefit of communities who may rely on their good intentions. For this reason, what is needed in LMICs, but also globally, is the introduction of rules, regulations and clear systems of accountability imposed by governments and international institutions

such as the WHO. The involvement of international bodies is critical. Countries with limited resources and great healthcare needs could have limited bargaining power against powerful global companies, which are able to interfere with national health policies.²⁰ Importantly, it should not be left to individuals on the ground to ‘negotiate’ their relationships with companies looking to harvest their data. This is not to say that all autonomy should be taken away from individuals about who access their data and for what reason. Rather, the framework within which these negotiations happen need to be constructed at a higher level, and with the public and global good in mind. By developing and applying a regulatory framework that sets the boundaries of operation and engagement, and ensures that the rights and benefits of people on the ground are served and protected is more likely to lead to a situation where appropriate and beneficial AI is developed and deployed.

Conclusion

AI has the potential to transform the way that healthcare is accessed and delivered around the world. Populations in LMICs also stand to benefit from these technological advancements. Even in clinical fields such as surgery which require access to professional and technological expertise, AI could help bring these expertise closer to remote populations, who would not have had access to it otherwise. Yet, the ability of AI to deliver on its promises depends on successfully resolving the ethical and practical issues identified, including that of explainability and algorithmic bias. Even though such issues might appear as being merely practical or technical ones, their closer examination uncovers questions of value, fairness and trust. It should not be left to AI developers, being research institutions or global tech companies, to decide how to resolve these ethical questions. Particularly, relying only on the trustworthiness of companies and institutions to address ethical issues relating to justice, fairness and health equality would be unsuitable and unwise. The pathway to a fair, appropriate and relevant AI necessitates the development, and critically, successful implementation of national and international rules and regulations that define the parameters and set the boundaries of operation and engagement. Establishing these rules and regulations is a precondition for ensuring that AI will be fit for the purpose of serving the public and global good.

¹ WHO. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.

² Hadley T. D., Pettit R. W., Malik T., Khoei A. A., Salihu H. M., Artificial Intelligence in Global Health -A Framework and Strategy for Adoption and Sustainability. *International Journal of Maternal Child Health and AIDS*. 2020; 9(1):121-127. doi: 10.21106/ijma.296. Epub 2020 Feb 10. PMID: 32123635; PMCID: PMC7031870.

³ Hashimoto, Daniel A et al. Artificial Intelligence in Surgery: Promises and Perils. *Annals of surgery* 2018, 268(1): 70-76. doi:10.1097/SLA.0000000000002693

⁴ Wahl B, Cossy-Gantner A, Germann S, Schwalbe, N. Artificial intelligence (AI) and global health: how can AI contribute to health in resource poor settings? *BMJ Glob Health* 2018;3:e000798. doi:10.1136/bmjgh-2018-000798

⁵ Kerasidou, A., Artificial Intelligence and the ongoing need for empathy, compassion and trust in healthcare, *Bulletin of the World Health Organisation* 2020; 98:245-250. doi: <http://dx.doi.org/10.2471/BLT.19.237198>

⁶ Grote, T, and Philipp B. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics* 2020; 46(3): 205-211. doi:10.1136/medethics-2019-105586

⁷ McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. *J Med Ethics*. 2019 45(3):156–60. doi: <http://dx.doi.org/10.1136/medethics-2018-105118> PMID: 304671

-
- ⁸ Dalton-Brown, S. The Ethics of Medical AI and the Physician-Patient Relationship. *Cambridge Quarterly of Healthcare Ethics* 2020; 29(1): 115-121. doi:10.1017/S0963180119000847
- ⁹ Amann, J., Blasimme, A., Vayena, E. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020; 20:310
<https://doi.org/10.1186/s12911-020-01332-6>
- ¹⁰ Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019; 366(6464):447-453. doi: 10.1126/science.aax2342. PMID: 31649194.
- ¹¹ WHO 2021
- ¹² Ibid.
- ¹³ Kerasidou, A. The role of trust in global health research collaborations', *Bioethics* 2018;
<https://doi.org/10.1111/bioe.12536>
- ¹⁴ Tindana, P, Molyneux, S, Bull, S, Parker, M. 'It is an entrustment': Broad consent for genomic research and biobanks in sub-Saharan Africa. *Developing World Bioeth.* 2019; 19: 9– 17.
<https://doi.org/10.1111/dewb.12178>
- ¹⁵ Guillemin M, Barnard E, Allen A, et al. Do Research Participants Trust Researchers or Their Institution? *Journal of Empirical Research on Human Research Ethics*. 2018;13(3):285-294. doi:10.1177/1556264618763253
- ¹⁶ Jao I, Kombe F, Mwalukore S, Bull S, Parker M, et al. Research Stakeholders' Views on Benefits and Challenges for Public Health Research Data Sharing in Kenya: The Importance of Trust and Social Relations. *PLOS ONE* 2015; 10(9): e0135545. <https://doi.org/10.1371/journal.pone.0135545>
- ¹⁷ Ipsos Mori. Public attitudes to commercial access to health data. 2016.
<https://wellcome.org/sites/default/files/public-attitudes-to-commercial-access-to-health-data-wellcome-mar16.pdf>
- ¹⁸ Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. 2020 May 16;395(10236):1579-1586. doi: 10.1016/S0140-6736(20)30226-9. PMID: 32416782; PMCID: PMC7255280.
- ¹⁹ Wright, S. Trust and Trustworthiness. *Philosophia* 2010; 38, 615–627
<https://doi.org/10.1007/s11406-009-9218-0>
- ²⁰ Hern, A. "Apple and Google block NHS COVID app update over privacy breaches" in *The Guardian*, Apr. 2021, [online] Available: <https://www.theguardian.com/world/2021/apr/12/apple-and-google-block-nhs-covid-app-update-over-privacy-breaches>.