An ethically mindful approach to AI for health care

Health-care systems worldwide face increasing demand, a rise in chronic disease, and resource constraints.[1] At the same time, the use of digital health technologies in all care settings has led to an expansion of data. These data, if harnessed appropriately, could enable health-care providers to target the causes of ill-health and monitor the effectiveness of preventions and interventions. For this reason, policy makers, politicians, clinical entrepreneurs, and computer and data scientists argue that a key part of health-care solutions will be artificial Intelligence (AI), particularly machine learning.[2,3] AI forms a key part of the National Health Service (NHS) Long-Term Plan (2019) in England, the US National Institutes of Health Strategic Plan for Data Science (2018), and China's Healthy China 2030 strategy (2016).

The willingness to embrace the potential future of medical care, expressed in these national strategies, is a positive development. Health-care providers should, however, be mindful of the risks that arise from AI's ability to change the intrinsic nature of how health care is delivered. Such potential AI transformations raise ethical risks that are normative, epistemic, or overarching. Such risks relate to inconclusive, inscrutable, or misguided evidence; unfair outcomes or transformative effects; or traceability.[4] Examples of these risks are described in the table.[4]

To mitigate these risks, a bold and systematic approach is needed to the implementation of AI solutions in health care that recognises the challenges and addresses them directly. Crucially, this approach must not rely solely on hard governance measures, such as new statutory obligations, that are designed in response to calls for the development of a robust regulatory system.[8,9] These measures are necessary but insufficient. Regulations provide only the necessary rules of the game, not the best strategy to win it. What is needed is an ethical focus on the end user, and their expectations, demands, needs, and rights.[10] The challenge is the insufficiently consistent approach to this kind of analysis.

To date, responses to the ethical risks, such as the NHS Code of Conduct for Data-Driven Health and Care Technology[11] and other non-sector specific ethical codes,[12] centre on protecting the individual. This approach is understandable because this is the

	Explanation	Medical example
Epistemic concerns		
Inconclusive evidence	Algorithmic conclusions are probabilities that are not infallible; they are rarely sufficient to posit the existence of a causal relationship	EKG readers in smartwatches may "diagnose" a patient as having arrhythmia when it may be due to a fault with the watch not being able to accurately read that user's heartbeat (eg, due to the colour of their skin) or the "norm" is inappropriately calibrated for that individual
Inscrutable evidence	Receivers of an algorithmic decision very rarely have full oversight of the data used to train or test an algorithm or the datapoints used to reach a specific decision	A clinical decision support system deployed in a hospital may make a treatment recommendation, but it may not be clear on what basis it has made that "decision", raising the risk that it has used data that are inappropriate for the individual in question or that there is a bug in the system leading to issues with over or under prescribing[5]
Misguided evidence	Conclusions can only be as reliable (but also as neutral) as the data they are based on	Use of image recognition algorithms trained on a dataset primarily composed of white patient images being used in Asia led to issues with accuracy of results[6]
Normative concerns		
Unfair outcomes	An action can be found to have more of an impact (positive or negative) on one group of people	An algorithm "learns" to prioritise patients it predicts to have better outcomes for a particular disease; this turns out to have a discriminatory effect on people within the black and minority ethnic communities
Transformative effects	Algorithmic activities, such as profiling, reconceptualise reality in unexpected ways	An individual using a personal health app has limited oversight over what passive data it is collecting and how those are being transformed into a recommendation to improve, limiting their ability to challenge any recommendations made and a loss of personal autonomy and data privacy[7]
Overarching		
Traceability	Harm caused by algorithmic activity is hard to debug (to detect the harm and find its cause), and it is hard to identify who should be held responsible for the harm caused	If a decision made by clinical decision support software leads to a negative outcome for the individual, it is unclear who to assign the responsibility, or liability to and therefore to prevent it from happening again

Table: **Epistemic, normative, and overarching ethical concerns related to algorithmic use in health care[4]**

level of analysis on which existing literature focuses, as illustrated in the table. However, the epistemic, normative, and overarching ethical risks associated with AI use in health care also arise at the relationship, group, institutional, and societal levels. For example, epistemic concerns related to misguided evidence and AI triaging chatbots, diagnostic image recognition systems, or clinical decision support software could cause harm to individuals, if the algorithm misjudges the severity of an individual's symptoms; patient–clinician relationships, if the patient trusts diagnosis of the algorithmic solution more than that of their clinician; groups and society, if biased training datasets lead to disproportionately better or worse health outcomes for different groups of people; and health-care institutions, if the inscrutability of the algorithm leads to patient safety issues that regulators are unable to address properly, resulting in loss of public trust. Thus, any ethical analysis of an AI system by health-care governing bodies must consider how potential ethical harms arise at different levels of analysis and at different stages of an algorithm's lifecycle.

This stage-by-stage analysis is essential since the ethical impact of an algorithm can be altered in either direction at any stage of development, sometimes recursively. For example, any pathology-recognising algorithm designed pro-ethically and independently validated could still cause harm if deployed in a health-care system unable to cope with the associated increase in patient volume. The ability of the system to operate effectively and safely would be impacted and might lead to individuals living with a potentially worrying diagnosis, with no help, for longer than before the algorithmic system was deployed. This is why ethical review cannot be a one-off, tick-box exercise and must be repeated regularly in a consistent way.

An internationally standardised and structured ethical review guideline should be developed by an advisory committee made up of technologists, clinicians, bioethicists and data ethicists, lawyers, human rights experts, and patient representatives. Development of the guideline should follow the best practice approaches set out by the Guideline International Network[13] so that it can be added to the International Guideline Library and be adopted by health-care systems looking to implement AI. The ultimate ambition, for those developing this guideline, should be to ensure that no AI solution is procured and implemented until a thorough ethical analysis of it has been completed according to the guideline, and a mechanism is established for the regular review of this analysis. Passing this evaluation should be a requirement even if the solution in question is, technically speaking, legally compliant with existing regulations. Only once this is the case will health-care systems be able to enjoy the dual advantage of ethical AI by capitalising on the opportunities while appropriately and proactively mitigating the risks to achieve the best possible outcomes for all.[14]

For the **International Guideline Library** see https://g-i-n.net/library/international-guidelines-library

JM is an employee of NHSX; NHSX had no oversight or influence over the writing of this Comment. We declare no other competing interests.

Jessica Morley, Luciano Floridi
jessica.morley@kellogg.ox.ac.uk

Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK (JM, LF); and Alan Turing Institute, London, UK (LF)

1 Martin GP, Sutton E, Willars J, Dixon-Woods M. Frameworks for change in healthcare organisations: a formative evaluation of the NHS Change Model. *Health Serv Manage Res* 2013; **26:** 65–75.
2 Chin-Yee B, Upshur R. Three problems with big data and artificial intelligence in medicine. *Perspect Biol Med* 2019; **62:** 237–56.
3 Harerimana G, Jang B, Kim JW, Park HK. Health big data analytics: a technology survey. *IEEE Access* 2018; **6:** 65661–78.
4 Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. *Big Data Soc* 2016; **3:** 205395171667967.
5 Wachter RM. The digital doctor: hope, hype, and harm at the dawn of medicine's computer age. New York: McGraw-Hill Education, 2015.
6 Liu C, Liu X, Wu F, Xie M, Feng Y, Hu C. Using artificial intelligence (Watson for Oncology) for treatment recommendations amongst Chinese patients with lung cancer: feasibility study. *J Med Internet Res* 2018; **20:** e11087
7 Kleinpeter E. Four ethical issues of "e-health". *IRBM* 2017; **38:** 245–49.
8 Challen R. Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019; **28:** 231–37.
9 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; **25:** 30–36.
10 Floridi L. Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philos Transact A Math Phys Eng Sci* 2018; **376:** 20180081.
11 Department of Health and Social Care. Guidance Code of conduct for data-driven health and care technology. July 18, 2019. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology (accessed Dec 10, 2019).
12 Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Machine Intell* 2019; **1:** 389–99.
13 Qaseem A. Guidelines International Network: toward international standards for clinical practice guidelines. *Ann Intern Med* 2012; **156:** 525.
14 Floridi L, Josh Cowls J, Beltrametti M, et al. AI4People—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds Mach* 2018; **28:** 689–707.