BMJ Health & Care Informatics

Can medical algorithms be fair? Three ethical quandaries and one dilemma

Kristine Bærøe ,[1] Torbjørn Gundersen,[2] Edmund Henden,[2] Kjetil Rommetveit[3]

To cite: Bærøe K, Gundersen T, Henden E, *et al*. Can medical algorithms be fair? Three ethical quandaries and one dilemma. *BMJ Health Care Inform* 2022;**29**:e100445. doi:10.1136/bmjhci-2021-100445

Received 11 July 2021
Accepted 10 December 2021

[1]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway
[2]Centre for the Study of Professions, Oslo Metropolitan University, Oslo, Akershus, Norway
[3]Center for the Study of the Sciences and Humanities, University of Bergen, Bergen, Hordaland, Norway

Correspondence to
Dr Kristine Bærøe;
kristine.baroe@uib.no

ABSTRACT

Objective To demonstrate what it takes to reconcile the idea of fairness in medical algorithms and machine learning (ML) with the broader discourse of fairness and health equality in health research.

Method The methodological approach used in this paper is theoretical and ethical analysis.

Result We show that the question of ensuring comprehensive ML fairness is interrelated to three quandaries and one dilemma.

Discussion As fairness in ML depends on a nexus of inherent justice and fairness concerns embedded in health research, a comprehensive conceptualisation is called for to make the notion useful.

Conclusion This paper demonstrates that more analytical work is needed to conceptualise fairness in ML so it adequately reflects the complexity of justice and fairness concerns within the field of health research.

Summary

What is already known?

- Biases in data, modelling and human review impact the fairness of the outcome of machine learning (ML).
- ML fairness in healthcare involves the absence of prejudices and favouritism towards an individual or group based on inherent or acquired characteristics, while fairness in health more broadly understood addresses historical fundamental socioeconomic biases that create health inequality within populations.
- Healthcare systems can fairly mitigate unjust health inequalities by offering equal opportunities for healthy lives.

What does this paper add?

- This paper argues that ML fairness in healthcare depends on equal access to healthcare systems.
- It demonstrates how a full conceptualisation of ML fairness in health is conditioned by a complex nexus of different fairness concerns.
- It calls for a reconceptualisation of ML fairness in health that acknowledges this complexity.

INTRODUCTION

Machine learning (ML) refers to algorithms that improve their performance independent of human designers. Several biases are involved in developing and applying ML, such as data biases (eg, historical and representation biases), modelling/design biases (eg, evaluation and aggregation biases) and human review biases (behavioural and social biases).[1 2] Biases affect the fairness of the ML process's outcome and deployment by wrongly skewing the outcome. 'Fairness' can be understood in different ways, but is usefully defined in the context of ML-based decision making as 'the absence of any prejudice or favouritism towards an individual or a group based on their inherent or acquired characteristics'.[3] Thus, when operationalising fairness into ML systems applied in health, the goal should be to eradicate biases in the processes of data sampling, modelling and human review so that the ML process does not promote health advantages or disadvantages for any individuals or groups based on their inherent or acquired characteristics. Assessing what kind of characteristics are assumed relevant for a fairness approach relies on normative ideas about justice. In healthcare, the World Medical Association's Declaration of Geneva identifies 'age, disease or disability, creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor' as examples of factors that should not impact the doctors' duty towards their patients.[4] Thus, ML failing to perform adequately to certain patients with such characteristics can be judged unfair.

In parallel, a more comprehensive and ambitious conceptualisation of fairness in health is discussed in the literature addressing how to distribute healthcare justly. Fairness is here understood in terms of how healthcare needs are unequally distributed within and across populations in the first place, which calls for justly allocated healthcare to reduce historically and socially conditioned inequalities. Theoretically, this aim is captured by egalitarian approaches to ensure, for example, equal opportunities[5] or capabilities,[6] or social justice.[7 8] Politically, it is reflected in empirically informed reports on observed health inequalities (eg, WHO's report on closing the

gap of inequalities in a generation[9] and in the Sustainable Development Goal of promoting health equality[10]). In clinical settings, work has been carried out to clarify the appropriateness of considering socioeconomic factors to circumvent their adverse impact on patients' ability to benefit from treatment.[11] This work has been translated into a call for revising and clarifying the way 'social standing' requires clinical attention in the World Medical Association's Declaration of Geneva.[12]

Unjust health inequality is influenced by inequality in the socioeconomical, cultural and environmental factors (eg, access to clean water) that shape people's living conditions.[13] Although theories diverge as to what makes the resulting health disparities unfair, there is broad consensus that health inequality associated with socioeconomic determinants of health creates inequity and calls for amendment.[14] For this reason, ML fairness should not only be about avoiding prejudices and favouritism, but also about reducing unfair health inequalities,[15] particularly those associated with socioeconomic health determinants . In line with Rajkomar and colleagues' reasoning,[15] to avoid ML in healthcare contributing to maintaining or reinforcing health inequities, fairness should be operationalised into ML processes by ensuring equal outcome across socioeconomic status, equal performance of models across socioeconomic groups, as well as equal allocation of resources.

Against this backdrop, this paper aims to answer the following question: How can the *narrow* fairness discourse related to ML and absence of prejudice and favouritism, and the *broader* fairness discourse related to unjust health equality be reconciled in a comprehensive conceptualisation of ML fairness that can be operationalised to prevent health inequity from being maintained or reinforced by healthcare systems? A more comprehensive notion of fairness in ML healthcare can be used to articulate commitments of fairness and help structure guidelines and recommendations.[16]

We start the discussion by clarifying the nature of the ML algorithms we focus on and present two distinct versions of 'justice' (substantive and procedural). We then argue that an adequate notion of ML fairness depends on a comprehensive approach to *fair access* to healthcare, which is inherently connected with other fairness challenges calling for practical solutions. Next, we identify and describe three interrelated fairness quandaries and one fairness dilemma related to obtaining ML fairness in health. A meaningful conceptualisation of ML fairness, which can be implemented to avoid inequitable patient outcomes, must reflect this complex, intertangled nexus of fairness concerns.

METHOD

The methodological approach used in this paper is theoretical and ethical analysis.

RESULTS

By applying this method, we identify three ethical quandaries and a dilemma related to ML fairness in healthcare. First, there is what we call 'the unfair data quandary'. Second, there is 'the unfair design quandary'. Third, there is 'the reasonable disagreement quandary'. Finally, there is the dilemma that arises from trade-offs between fairness and accountability. Figure 1 illustrates our approach.

DISCUSSION

ML refers to algorithms that improve their performance based on previous results independently of human designers. An important subset of ML with much promise in medicine is deep learning algorithms, which process inputs (eg, data such as pictures, videos, speech and text) to provide output such as identified patterns,

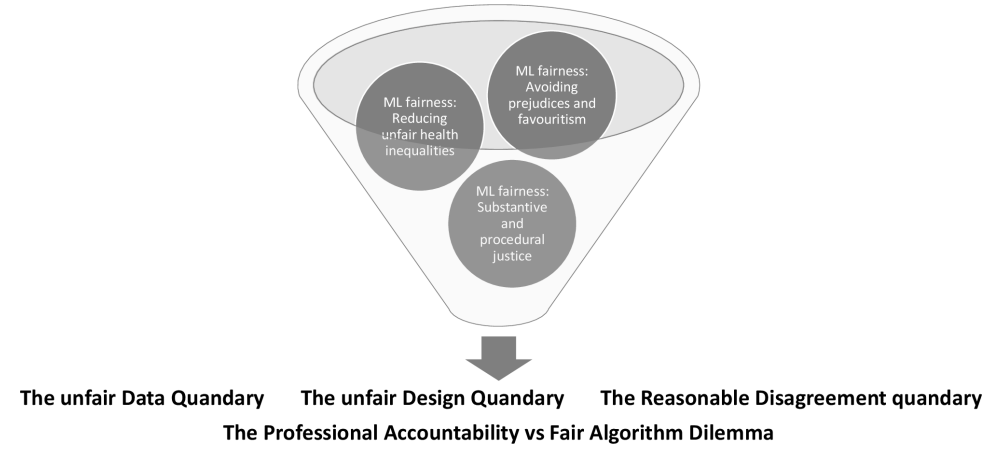

Figure 1 Three interrelated fairness quandaries and one fairness dilemma related to obtaining machine learning (ML) fairness in health are identified in this ethical analysis. A meaningful conceptualisation of ML fairness, which can be implemented to avoid inequitable patient outcomes, must reflect this complex, intertangled nexus of fairness concerns. This figure is made by the first author.

classifications or predictions.[17] Analogous to the animal brain, the mechanisms in deep learning are 'deep neural networks' consisting of hierarchically structured layers of 'neurons'. To work effectively, the neural networks are trained on vast data sets, which are sometimes labelled by humans (as in supervised learning) or they identify patterns in data sets on their own (as in unsupervised learning).[18] Due to its ability to identify pattern in vast data sets much faster and often more accurately than medical doctors, health professionals and scientists are able to, ML algorithms have the potential to make the detection, prediction and treatment of disease more effective.[17 19]

Substantial and procedural justice

Fairness can be analysed in terms of distinct principles of what justice requires (*substantive justice*) or in terms of the acceptability of how the decision is made (*procedural justice*).[20] The assumption behind procedural justice is that even though there may be widespread disagreement about what it would be just to do (eg, how to prioritise healthcare with resource scarcity), the affected parties may be expected to agree on what conditions must be in place to make the decision-making process fair.[21] Procedural justice requires, for example, that affected parties are treated equally by considering all interests at stake, and that decisions are based on reasons that individuals can recognise as relevant and reasonable.[21] Both versions of justice are relevant for fair decision making and how fairness issues come into play in relation to ML fairness.

Three quandaries and one dilemma

The unfair data quandary

The first quandary is related to biased data. This quandary states that groups of people not accurately represented in the training data of ML algorithms could receive diagnosis and treatment recommendations systematically imprecise in their disfavour. Since healthcare data typically emerge from contact with and/or use of the healthcare system, the extent to which people have *access* to healthcare will predict their inclusion in ML training data.

A conceptualisation of 'access to healthcare' can be divided into a supply side of the organised service and a demand side of patients' ability to benefit from the organised care. 'Access to healthcare' can be conceptualised across different phases involved in having a healthcare need met, that is, having a need, perceiving a need and desire for care, seeking healthcare, reaching healthcare services, using healthcare services and obtaining healthcare outcomes.[22] This broad approach to access to healthcare is useful for a nuanced investigation of where, when, how and by whom inequality in access can emerge under the impact of organised healthcare itself.

Healthcare services can uphold or reinforce social inequality in health if access to services requires capacities associated with socioeconomic conditions unequally distributed in the population. If the supply side is not carefully developed to meet the social and economic challenges related to people's abilities to reach and obtain care (eg, ability to pay or follow prescribed regimes, understanding of their own health or how the system works, cultural conflicts), data gathered from these services could be biased favouring those with the abilities to overcome barriers (eg, by paying for health insurance). Thus, unequal access skews the representativeness of big data gathered within the healthcare system to the advantage of those who have historically been able to use it. As this is the available data that ML algorithms are trained on, the detection of disease and clinical recommendations might not be equally apt for the groups that experience barriers in reaching, receiving and benefiting from care. This can be so if these latter groups overlap with relevant biological differences related to ethnical background, or if lifestyle issues related to socioeconomic challenges, impact the uptake of treatments. For the training data to be fair, the real-world conditions for access to healthcare must be equal in the sense that socioeconomic barriers do not prevent people from obtaining care.

The challenge to ensure fairness stemming from a lack of representative data is structural. Use of historically biased data combined with underdeveloped labelling creates racial biases in healthcare management of populations.[23 24] Space does not allow us to do justice to the vast literature on algorithm fairness and suggestions to mitigate algorithm biases. For a comprehensive overview, there is a framework proposed by Suresh and Guttag, which identifies the multiple sources of downstream harms caused by ML through data generation, model building, evaluation and data deployment, and also describes mitigation techniques for targeting the same sources.[2] As noted above, there are strong ethical and political calls to promote equal access to high-quality healthcare for all. To avoid a situation where ML unfairly maintains (or even reinforces) inequality in health outcomes, coordinated initiatives could be directed comprehensively towards identifying barriers and seeking innovative solutions to promote equal access to healthcare in the first place along all dimensions of supplying and demanding healthcare. Developers of ML systems, ethicists and funding bodies could join forces and gear attention towards mitigating the structural unfairness of unequal access to healthcare before addressing the inequitable outcome of this unfairness.

The unfair design quandary

Let us assume that comprehensive work has been done to ensure equal access to healthcare for all, which can enable fairness in algorithms deployed at the point of care. Now fairness is an issue about what kind of ML-based healthcare ought to be developed, that is, what kind of ML should be prioritised. How should this fairness aspect be ensured in the *design phase* when fair design then ideally must include broad oversight of consequences and justified priority-setting decisions *before* ML interventions have been developed and tested?

First, the ethical issues that arise from an ML system will depend on its practical application and purpose: is the system used for home monitoring, clinical decision support, improved efficiency and precision in testing, distribution and management of medicines, or something else? What kind of disease or ailment is being addressed? The ethical problems will be different and include different actors.

Next, one should ask: is ML needed, or do existing approaches work better? This is about the performance of ML, for example in terms of improved prediction.[25] But it is also about getting the process of interpreting what it means 'to work better', right. Who should decide that?

Depending on the problem being addressed, different actors will be involved. Design is not a linear process.[26] It depends on reiterated cycles of design, implementation, testing (including with other data), assessment and evaluation. This is even more so with ML algorithms, as they might display unpredictable outcomes (depending on input data, but also coding and algorithms). They therefore need constant human monitoring and assessment. The European Commission, for example, emphasise the need for stakeholder involvement throughout all the cycles.[27]

Potentially, these phases will involve inputs from people such as medical doctors, nurses, hospital administrators, health economists, other technical people and (ideally) the patients themselves. This then poses the question of the competencies that should enter into the design, implementation and testing phases, how they should be made to cooperate, and what kinds of expertise should count. Whose professional perspective may frame the initial understanding of the problem, what happens to dissenting voices, and what about patients' perspectives and autonomy? If these challenges to justice are not explicitly addressed, it might create an unfair design quandary . Procedural justice requires developing adequate and fair decision-making institutions for collaboration. This must be organised so all stakeholders can recognise them as being fair by including general requirements on transparency, reasonable justifications and opportunities for revision.[21] Still, figuring out how to best do so in these contexts requires more research.

The reasonable disagreement quandary

People are expected to disagree about principles of justice, what societal challenges one will trade off to improve people's health, and what opportunity costs one will accept to achieve health equality in the design process. How should such ethical disagreements be resolved? Procedural fairness addresses the moral equality of anyone involved in or being affected by a decision by arranging a decision-making process in a way that all can find acceptable. This means, for example, that all stakeholders must be included, allowing everyone to voice their concerns and listen to them, ensuring transparency of the rationales for the decision, and offering mechanisms to appeal.[21] In the case of designing and applying ML in medicine, there are multiple groups of experts, professions and other stakeholders that might play a central role in this kind of ethical deliberation, for instance medical doctors, nurses, hospital administrators, health economists, technicians, ML developers, patients, and the public in general. However, such inclusive deliberation might not be feasible to arrange every time an ML system is developed. The complex task of identifying, understanding and weighing all relevant medical, ethical, economical and societal issues to consider and reasonably justify what to prioritise in order to apply ML requires substantial technical and disciplinary expertise. Also, to ensure the prioritisation is adequately reflected in the design process, it is crucial to rely on ML experts and their interpretations when translating normative decisions into algorithms. This 'reasonable disagreement quandary' requires a fix in terms of fairness, but an overall fair decision-making process can be difficult to realise. Moreover, the fairness of leaving the decision to trained decision makers or technical experts and the substantial principles of justice they happen to hold, is also questionable.

More research is required to learn how to better maximise inclusiveness and transparency and monitor whether ethical and political prioritisations are captured in ML systems in a meaningful way. The aim should be to accommodate procedural fairness. However, realism is needed in identifying and articulating the limitations of such a fairness approach. A hybrid model of fairness based on substantial and procedural justice might emerge as a solution.

A final ethical dilemma

Let us, for the sake of argument, assume that the above quandaries are solved. Let us suppose that measures have been taken to ensure that the training data are not skewed, that adequate institutional conditions for collaboration between stakeholders and designers have been established, and that an acceptable model of procedural fairness has been developed. There is still, however, the following dilemma that needs to be addressed: while medical algorithms might improve fairness by eliminating biases that otherwise might affect the decisions of healthcare professionals and therefore result in more equitable access to healthcare services, they might also reduce the accountability of healthcare professionals for these same decisions. Algorithmic decision systems are built so it makes it difficult to determine why they do what they do or how they work. For example, neural networks that implement deep learning algorithms are large arrays of simple units, densely interconnected by very many links. During training, the networks adjust the weights of these links to improve performance, essentially deriving their own method of decision making when trained on a decision task. They therefore run independently of human control and do not necessarily provide an interpretable representation of what

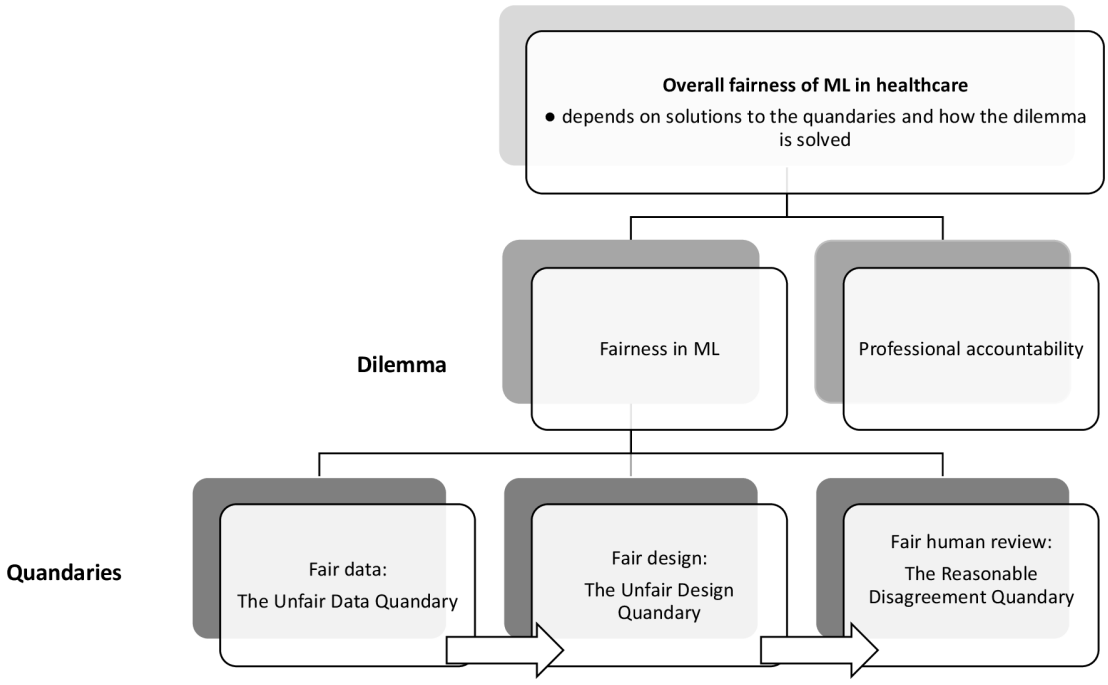

Figure 2 Overall fairness of machine learning (ML) in healthcare depends on solutions to the three interrelated quandaries and the dilemma. This figure is made by the first author.

they do.[28 29] The problem with this is that *professional accountability* cannot be enforced without explainability. Professional accountability implies that it is justifiable to ask a healthcare professional to explain their actions and to clearly articulate and justify the decisions they have made. Providing such an explanation engenders trust in the process that led to the decision and confidence that the healthcare professional in charge of the process acted fairly and reasonably. It is true, as some have pointed out, that lack of explainability in medicine is not uncommon—sometimes it may be close to impossible to reconstruct the exact reasoning underlying the clinical judgement of a medical expert and there may be little knowledge of the causal mechanisms through which interventions work.[30] However, explainability is still important in some contexts, particularly in those requiring informed consent. In contexts where explainability *is* important, the potential opacity of ML algorithms suggests that some trade-off must be made between deferring to said algorithms (which might improve fairness but reduce accountability) and relying on human professional discretion (which might preserve accountability but increase the risk of biases). The dilemma is that neither option comes without ethical costs: either reduced accountability or (potentially) reduced fairness. Figure 2 shows how the quandaries and the dilemma are interrelated and part of a broad conceptualisation of ML fairness in healthcare.

CONCLUSION

We have demonstrated that operationalising fairness in ML algorithms in healthcare raises a whole host of fairness challenges across data, design and implementation biases, which all need to be solved before concluding that the algorithms are fair. Even if we have the ability to meet these challenges, we nevertheless face the problem of trading fair algorithms off against professional accountability. To avoid a rhetorical and insufficiently justified conception of fairness in ML technology, these fundamental and intangible challenges of fairness must be openly acknowledged and addressed. In addition, much more research on fair processes is called for to find ethically and politically sustainable responses to what fairness requires of ML algorithms employed in clinical care.

Acknowledgements The authors thank two anonymous reviewers for their valuable suggestions.

Contributors KB had the idea, drafted the first version and is acting as the guarantor. TG, EH and KR made substantive inputs to the draft and contributed to the revisions of the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD
Kristine Bærøe http://orcid.org/0000-0002-4626-7232

REFERENCES

1 Reagan M. Understanding bias and fairness in AI systems, 2021. Available: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3 [Accessed 03 Jul 2021].
2 Suresh H, Guttag JV. A framework for understanding unintended consequences of machine learning. *arXiv* 2019;2:190110002.
3 Mehrabi N, Morstatter F, Saxena N. A survey on bias and fairness in machine learning. *ArXiv* 2019:abs/1908.09635.
4 World Medical Association. Declaration of Geneva. Available: https://www.wma.net/policies-post/wma-declaration-of-geneva/2018 [Accessed 15 Jun 2020].
5 Daniels N. *Just health: meeting health needs fairly*. Cambridge University Press, 2007.
6 Sen A. Why health equity? *Health Econ* 2002;11:659–66.
7 Peter F. Health equity and social justice. *J Appl Philos* 2001;18:159–70.
8 Braveman PA, Kumanyika S, Fielding J, *et al*. Health disparities and health equity: the issue is justice. *Am J Public Health* 2011;101 Suppl 1:S149–55.
9 Marmot M, Friel S, Bell R, *et al*. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* 2008;372:1661–9.
10 United Nations. *Transforming our world: the 2030 agenda for sustainable development*. New York: United Nations, Department of Economic and Social Affairs, 2015.
11 Bærøe K, Bringedal B. Just health: on the conditions for acceptable and unacceptable priority settings with respect to patients' socioeconomic status. *J Med Ethics* 2011;37:526–9.
12 Bringedal B, Bærøe K, Feiring E. Social Disparities in Health and the Physician's Role: A Call for Clarifying the Professional Ethical Code. *World Medical Journal* 2011;5:196–8.
13 Dahlgren G, Whitehead M. Policies and strategies to promote social equity in health. Background document to WHO - Strategy paper for Europe. *Arbetsrapport* 1991.
14 Wester G, Bærøe K, Norheim OF. Towards theoretically robust evidence on health equity: a systematic approach to contextualising equity-relevant randomised controlled trials. *J Med Ethics* 2019;45:54–9.
15 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
16 Wawira Gichoya J, McCoy LG, Celi LA, *et al*. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021;28:e100289.
17 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
18 Franklin S. History, motivations, and core themes. In: *The Cambridge Handbook of artificial intelligence*, 2014: 15–33.
19 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med Overseas Ed* 2019;380:1347–58.
20 Miller D. Justice. In: Zalta EN, ed. *The Stanford encyclopedia of philosophy*. 2017 Edition, 2021. https://plato.stanford.edu/archives/fall2017/entries/justice/
21 Daniels N, Sabin J. *Setting limits fairly: can we learn to share medical resources?* Oxford University Press, 2002.
22 Levesque J-F, Harris MF, Russell G. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. *Int J Equity Health* 2013;12:18.
23 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
24 Benjamin R. Assessing risk, automating racism. *Science* 2019;366:421–2.
25 Desai RJ, Wang SV, Vaduganathan M, *et al*. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open* 2020;3:e1918962.
26 Stewart J, Williams R. 10. The wrong trousers? Beyond the design fallacy: social learning and the user. In: Rohracher H, ed. *User involvement in innovation processes strategies and limitations from a socio--technical perspective*. Munich: Profil Verlag, 2005.
27 Independent High-Level Expert Group on Artificial Intelligence (AI IHLEG). *Ethics guidelines for trustworthy AI*. Brussels: European Commission, 2019.
28 Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc* 2016;3:205395171562251–12.
29 de Fine Licht K, de Fine Licht J. Artificial intelligence, transparency, and public decision-making. *AI Soc* 2020;35:917–26.
30 London AJ, Intelligence A. Artificial intelligence and black-box medical decisions: accuracy versus Explainability. *Hastings Cent Rep* 2019;49:15–21.