Cambridge Quarterly of Healthcare Ethics (2024), 33: 3, 311–322
doi:10.1017/S0963180123000257

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Hammer or Measuring Tape? Artificial Intelligence and Justice in Healthcare

Jan-Hendrik Heinrichs

Institute for Neuroscience and Medicine 7: Brain and Behaviour, Forschungszentrum Jülich GmbH, Jülich, Germany; Institute of Philosophy, RWTH Aachen University, Aachen, Germany
Email: j.heinrichs@fz-juelich.de

Abstract

Artificial intelligence (AI) is a powerful tool for several healthcare tasks. AI tools are suited to optimize predictive models in medicine. Ethical debates about AI's extension of the predictive power of medical models suggest a need to adapt core principles of medical ethics. This article demonstrates that a popular interpretation of the principle of justice in healthcare needs amendment given the effect of AI on decision-making. The procedural approach to justice, exemplified with Norman Daniels and James Sabin's *accountability for reasonableness* conception, needs amendment because, as research into algorithmic fairness shows, it is insufficiently sensitive to differential effects of seemingly just principles on different groups of people. The same line of research generates methods to quantify differential effects and make them amenable for correction. Thus, what is needed to improve the principle of justice is a combination of procedures for selecting just criteria and principles and the use of algorithmic tools to measure the real impact these criteria and principles have. In this article, the author shows that algorithmic tools do not merely raise issues of justice but can also be used in their mitigation by informing us about the real effects certain distributional principles and criteria would create.

Keywords: Artificial intelligence; justice; algorithmic fairness; accountability for reasonableness; algorithmic epistemic tools

Introduction to AI and Justice in Healthcare

Artificial intelligence (AI) is a new and powerful tool for a broad range of healthcare tasks, which ties in with previous developments in digital healthcare such as P4 (predictive, preventive, personalized, and participatory) systems medicine. P4 systems medicine gathers and integrates patient information across multiple areas—from genomics, transcriptomics, proteomics, metabolomics, to traditional anamnestic methods, to social and environmental influences into one or several predictive models.[1] With the inclusion of modern machine-learning algorithms, such models can predict—amongst others—health outcomes of a large set of possible interventions on all—or most—of these levels.

AI tools in healthcare are particularly suited to optimize the predictive part of P4 systems medicine. They have not only been used to improve the processing of *medical* data in support of diagnoses, of therapy decisions, or of hospital management.[2] Also, including *nonmedical* data in the processing with AI algorithms has extended their predictive power. One prominent example is the inclusion of social media data into the prediction and even diagnoses of mental disorders.[3]

The employment of such predictive models has received a mixed reception in the clinical community as well as in medical ethics. On the one hand, it promises clear advantages for clinical management and patient care. On the other hand, several risks have been identified, such as the risk of over-automating the clinical processes,[4] or of making extreme breaches of privacy possible.[5] Another issue that has received significant attention is that of medicalization[6] by the inclusion of more types of information into medical prediction and thereby into medical risk management. In addition, ethical debates about AI's extension

of the predictive power of medical modeling have suggested the necessity to amend established principles of medical ethics. In particular, Sebastian Laacke et al.[7] argue that the principle of autonomy, as introduced by Tom Beauchamp and James Childress,[8] needs to be amended by an opt-in requirement they introduce under the terms of "availability of alternatives" and "independence."

In this paper, I demonstrate that the contemporary interpretation of a further principle of medical ethics needs amendment, given the effect of AI on decision-making in healthcare contexts: the principle of justice. There is an increasing number of studies that show that using machine-learning systems in healthcare results in differential treatment for people belonging to different socioeconomic groups, ethnic backgrounds, gender, and so forth.[9,10,11,12,13] In this situation, one would expect that it is possible to turn to the principles of medical ethics, in particular to the principle of justice, to find means of evaluating and, where required, of mitigating or correcting this differential treatment. While this is, in principle, possible, the following will argue that the current interpretation of the principle of justice needs amendment before it can solve the issues raised by AI tools in healthcare.

One major challenge for theories of justice in healthcare is the number and diversity of decision problems. It comprises the more obvious question of resource allocation as well as questions of epistemic justice in factual, medical questions, the budgeting of different areas of healthcare from the provision of medical services over drug development and testing, education of practitioners, hospital, and private practice infrastructure, and many more. Even in the seemingly homogeneous decision problem of allocating scarce therapeutic resources to patients, there is surprising diversity. Different reasons can be brought forward in questions of organ allocation and of allocation of pain medication, for example. The current ethical solution to this diversity of decision problems, as well as to the lack of consensus concerning a substantive theory of justice, is a procedural approach. It is intended to provide general criteria under which particular decision processes are just.

It will be argued that this procedural approach as exemplified by Norman Daniels and James Sabin's *accountability for reasonableness* conception needs to be amended because, as research into algorithmic fairness has shown, it is insufficiently sensitive to the differential effects that seemingly just principles have on different groups of people. The same line of research has, however, generated methods to measure and quantify such differential effects and thereby make them amenable for evaluation and correction. Thus, what is needed to counter the problems of unfairness in AI algorithms is a combination of procedures for selecting just criteria and principles and methods to measure the real impact these criteria and principles have. This requirement for adequate measurements of the impact of criteria and principles is an amendment to the principle of justice as it is interpreted today. The latter will most likely comprise AI methods themselves.

AI Tools Raising Issues of Justice in Healthcare

An impression of how the introduction of AI-powered predictive models in healthcare has raised issues of justice can be gained from two current and one still fictional examples: (1) NarxCare is a prescription drug monitoring program (PDMP), originally intended to detect risks of drug misuse, diversion, and overdose.[14] As such, it is originally not intended as a system for healthcare but for law enforcement. NarxCare predicted forms of misuse where people visited several different medical practitioners and pharmacies, traveled long ways for that purpose, or paid for their medications in cash. It is unknown whether the system learned these criteria from the data or whether they have been preprogrammed. Against its original purpose, practitioners started to consult NarxCare in their therapeutic decisions, that is, for prescription of pain medication. One effect of this use has been to exclude people from access to adequate medical care. In particular, it excludes people who live far from specialists, who need several different specialists, or who cannot afford the services of insurance and credit card companies. (2) Ziad Obermeyer et al. have analyzed an algorithm used to assign patients to high-risk care programs, that is, healthcare programs suited to individuals with high healthcare risks.[15] Fair analysis shows how the choice of variables used as proxies for individual healthcare requirements can result in biased—in this case, racially biased—effects. The algorithm in question used past healthcare costs as a label to predict future healthcare requirements. But past healthcare costs can and do differ from past healthcare

requirements and thus are bad predictors for future requirements. In this case, it turned out that among people with the same past healthcare costs, people of color had significantly worse health than white individuals. When the authors used the individual's health status instead of previous healthcare costs, the number of people of color to be included in the higher-risk care increased significantly.

The moral issue—the nail sticking out—in these examples is not merely that they involve outcomes of algorithmic procedures that offend against our substantive intuitions about what justice requires. Rather, it is that differential effects are caused by the procedure being sensitive to information that was thought and intended not to play a role in the procedure in question. These two examples are characterized by the inclusion of nonmedical variables into their prediction. However, similar problematic effects can occur when biologically and medically relevant data are included in algorithmic analysis without exact control for its purpose.[16]

Here is a—still—fictional example: posttraumatic stress disorders (PTSDs) manifest differently in persons of different sex and gender; they have different physiological and anatomical manifestations, and there are also different gender-specific risk factors. As such, a more precise taxonomy and etiology of PTSD should take these variables into account. If one trained an algorithm to improve the taxonomy and etiology of mental illnesses, then such data should be included. At the same time, however, it is an empirical fact that individuals are exposed to traumatizing experiences to different degrees because of their gender.[17] Thus, social reaction to gender—and most likely not just to binary gender—plays a role in the prevalence of PTSD, confounding the information of gender-based influence on the disease. For a slightly different purpose, however, the relation is reversed. If one trained an algorithm for diagnostic purposes, then one might have to account for what is morally inacceptable: that most cultures' reaction to specific expressions of gender is not just a confounder but a risk factor for PTSD.

AI Tools Uncovering Issues of Justice in Healthcare

These examples seem, at first hand, to be amenable to a simple solution. We can treat, and should treat, justice in healthcare independently of and prior to questions of how to design our algorithmic tools. Once we are clear about what would be just allocations or procedures in healthcare, we can either put it into action manually or use algorithmic tools much like a hammer, that is, design and train algorithms to realize this measure of justice. This is indeed a common premise in several contributions in the debates about algorithmic fairness. However, this approach is not available anymore. I will argue that it is not available because algorithmic tools raise serious doubts about established ideas of what justice in healthcare requires. In order to explain this claim, the following will first sketch Daniels and Sabin's accountability for reasonableness conception as an example of the contemporary dominant procedural approach to justice in healthcare. Following up on a critique of this conception, I will show how algorithmic tools have revealed radical limitations to some of its components. However, far from being merely a danger to healthcare justice only, algorithmic tools can also be used to measure the real effects of principles of justice and thereby show paths to repair injustice, as I will consecutively discuss.

Justice: The Rawlsian Tradition

The Rawlsian tradition has strongly influenced contemporary conceptions of justice in healthcare. In particular, Daniels had a dominant influence on the field. Together with Sabin, he—and a number of other authors,[18,19] —has initiated a procedural turn in theorizing about healthcare justice.[20] This turn follows the tradition of the *later* John Rawls, focusing on political deliberative processes as (not: for) the solution to questions of justice (against seeing Daniels and Sabin as Rawlsians).[21] The principles and procedures of healthcare justice must be designed in a way that allows resolving reasonable disagreements about resource use. Their[22] procedural approach formulates four requirements for a fair decision procedure for allocative purposes:

1. *Publicity Condition*: Decisions regarding both direct and indirect limits to care and their rationales must be publicly accessible.
2. *Relevance Condition:* The rationales for limit-setting decisions should aim to provide a *reasonable* explanation of how the organization seeks to provide "value for money" in meeting the varied health needs of a defined population under reasonable resource constraints. Specifically, a rationale will be reasonable if it appeals to evidence, reasons, and principles that are accepted as relevant by fair-minded people who are disposed to finding mutually justifiable terms of cooperation.
3. *Revision and Appeals Condition:* There must be mechanisms for challenge and dispute resolution regarding limit-setting decisions, and, more broadly, opportunities for revision and improvement of policies in the light of new evidence or arguments.
4. *Regulative Condition:* There is either voluntary or public regulation of the process to ensure that conditions 1–3 are met.[23]

These four criteria are the core of what has come to be called the *accountability for reasonableness* approach. While there is ample critique of the details of this approach,[24,25,26,27,28,29,30] it has become one of the most important standards in the debate.

One of the most critical components of the *accountability for reasonableness* approach is the second, the relevance condition. In the following, it will be claimed that it cannot accommodate what we know today about the sensitivity of procedures to factors, which play no explicit role in them. Because this condition will be attacked in the following, it is only fair to describe its rationale beforehand.

The relevance condition is motivated by a core component of the social contract tradition, the idea that societal limitations to an individual's freedom have to be justified by reasons this person can—or must—accept.[31,32] Daniels and Sabin's formulation in this context is that "the reasons offered by decision makers must be those that persons affected by the decisions can recognize as relevant and appropriate."[33] In the further debate, it has been doubted whether the relevance condition is suited to capture the contractualist intuition.

Alex Friedman, for example, provides a detailed critique.[34] His initial observation is that, for some criteria, it is contested between reasonable people whether to include them in their decisions. Whether, for example, to include age in an allocation decision depends on the further choice to treat either all lives, all life-years, chances to a full life cycle or individual investments into lives as mattering equally.[35] All of these positions are reasonable and have been suggested in sophisticated ethical debates, but according to some, age should not play just any role in the deliberation about therapy allocation. Even if there should be consensus about which reasons to include, counter to Daniels and Sabin, it does not make the remaining decision less controversial. Disagreements about the weight of the different reasons are not necessarily less intractable or easier to solve. If inclusion into a healthcare program depends on the weight distribution in a particular consideration—such as efficiency versus need—it is to be expected that these weights can be as contested as those about the inclusion of reasons in the first place.[36]

Last but not least, Friedman—following a remark by Daniels and Sabin—points out that people will disagree about what reasonable people disagree about. For some, it will be completely unreasonable to disagree about the sanctity of all life and thus to even suggest excluding such a consideration from deliberations about healthcare, while others will insist that no reasonable person can expect them to consider *sanctity* of anything a reason.[37]

In another critical discussion, Gabriele Badano claims that Daniels and Sabin's relevance criterion goes against the Rawlsian tradition insofar as it allows a major role for cost-effectiveness analysis in considerations of justice. Cost-effectiveness analysis, however, starts out with a seriously wrong unit of concern, namely the utility generated by the goods, not with the interests of persons.[38] Thus, by allowing cost-effectiveness analysis, the *accountability for reasonableness* approach goes against the basic Rawlsian tenet, to base deliberation on the interest of individuals only, the principle of separateness of persons. Badano suggests replacing the overinclusive relevance condition with a stronger version from the contractualist tradition such "that decisions should be made according to principles that no one could reject in a situation in which everyone is committed to proposing principles that no other similarly

motivated person could reject."[39] He calls that the "full acceptability condition," trusting that it is suited to exclude aggregative considerations, and thus violations of the principle of separateness of persons, from just procedures. I take the above criticism, especially the call for a full acceptability condition, to suggest corrections and amendments to and not a rejection of *accountability for reasonableness.*

Algorithmic Bias Versus the Relevance Condition

It would, however, be futile to try to decide between the different means of accounting for the contractualist intuition just yet. First, an even more serious critique will have to be considered. The debate about algorithmic fairness has revealed a previously disregarded weakness in the relevance condition. The relevance condition insists that an institution's rationale for limit-setting decisions about meeting the health needs of a population "will be reasonable if it *appeals to* evidence, reasons, and principles that are accepted as relevant by fair-minded people who are disposed to finding mutually justifiable terms of cooperation."[40]

The debate about algorithmic fairness has set off with the insight that procedures need not process, that is, *appeal to,* specific information, which fair-minded people would not accept as relevant, to have the same effect as if they did.[41,42] In particular, contemporary learning algorithms tend to still be sensitive to information that is actively withheld from their processing. Other information will stand proxy for what has come to be called "protected attributes." The most common examples in the debate concern the attributes "ethnic background" and "gender." For several different reasons (bias in the data, bias in the measurement procedures, etc.),[43] learning algorithms often come to generate different results concerning people of different gender and ethnic background, even if gender and ethnic background are not available variables for the system.

This effect clearly turned up before the debate about algorithmic bias got off the ground, even in textbook examples, but these examples have been used to exemplify different issues. One such example has been discussed in early editions of the most influential book on biomedical ethics, Tom Beauchamp and James Childress' *Principles of Biomedical Ethics.*[44] They take it to be common ground that healthcare access should not depend on ethnic background. This demand is complicated to operationalize, as shown by the example of access to renal transplantation in the United States. The original policy that sought to "maximize the number of quality-adjusted life-years per transplanted organ"[45] resulted in a lower implantation rate for people of color because most organ donations come from white Americans. Tissue matching between donor and recipient affects the long-term survival of the transplant, and there are relevant differences in tissue matching between populations. Thus, even if ethnic background is not included in the decision procedure, and even if the decision procedure maximizes for correction of health-related loss of opportunity (quality-adjusted life-years), there is a clear differential effect on people of different ethnic backgrounds. A reformed policy, on the other hand, which gives less impact to tissue matching and thus results in less correction of health-related loss of opportunity, can correct for the original differential outcome by ethnic background. While, again, it does not include ethnic background as a relevant factor in individual decisions, it changes the allocation between whites and people of color to the advantage of the latter and thus makes healthcare access indirectly dependent on ethnic background.

Thus, even if fair-minded people agree that access to medical care should not depend on ethnic background, and even if allocation procedures are put into place, which do not involve ethnicity, differential treatment emerges for people of different ethnic groups. The debate about algorithmic fairness has resulted in many examples of this phenomenon, many related to ethnic background, gender, socioeconomic status, and so forth.

In many cases, the cause of this misalignment between the reasons and principles used in a given procedure and the effects of the procedure can be found in historic injustice. There are several reasons why historic injustice is continued or even exacerbated by what looks like just procedures and their principles. Amongst them is the "bias-in bias-out" problem of learning algorithms as well as the fact that equal treatment cannot be just given unequal starting conditions.

Annette Zimmermann and Chad Lee-Stronach[46] have recently defined procedural injustice thus: "On our view, an algorithmic decision procedure is procedurally unjust to individuals subject to it if and because the procedure fails to include relevant information about the effects of current and past substantive structural injustices, including—but not limited to—racial and gender injustice."[47] This formulation has clear weaknesses in that it defines one type of injustice (procedural) with another (structural) without providing any independent explication of the latter. Nevertheless, given the shortcomings of contemporary procedural approaches to justice, it would be adequate to drop the term "algorithmic" and the unclear reference to structural injustice in the definition, and insist that any decision procedure that fails to include relevant information is unjust. While this requirement does not speak as to which information is relevant, it seems clear that data concerning the effect of a decision process are forgoing their collection and use, and therefore seems to be unjust. Established decision procedures in many areas did exactly that; they did not include relevant information about the effects of possible decision structures in view of current and past social structures.

The debate about algorithmic fairness thus reveals that established procedural approaches to selecting criteria and principles of justice, in general, and the relevance condition of Daniels and Sabin's account, in particular, have a serious problem. The mere fact that certain reasons, principles, or criteria are not being used in a procedure does not suffice to make that procedure insensitive to them. Fair-minded people cannot rest assured that their procedures are just, simply because they *appeal to* the right kind of reasons and not to controversial reasons. They need to do more in order to work toward justice in healthcare. They need to make sure that their procedures are *sensitive* to the right criteria.

The Role of Algorithms in Just Procedures

If the argument presented so far is correct, then consequences might seem dire: Algorithmic tools in healthcare clearly raise significant issues of justice. The examples above made that much clear. At the same time, it turns out that the principle of justice in healthcare and the theory of justice we would typically use to cope with these issues—*accountability for reasonableness*—have shortcomings. Although these shortcomings came to general attention in the debate about algorithmic justice, they have been present under the surface for quite some time before.

One of the first things that this implies is that we cannot simply fall back on *accountability for reasonableness* to solve questions of algorithmic fairness. There might, however, be kernels of a solution in the attempt. A recent contribution by Pak-Hang Wong[48] has gone into this direction. He searches for means to go beyond the current state of the debate in algorithmic fairness and suggests using the *accountability for reasonableness* approach.

Wong motivates his argument with the observation that research into algorithmic fairness has shown that specific plausible technical measures of fairness regularly conflict when applied to the same algorithmic decision procedure. This observation has become famous in the aftermath of the complaint against the algorithmic tool for supporting parole decisions, COMPAS, which failed on one important fairness criterion (*separation*) because it was designed to comply with another (*sufficiency* or *calibration*).[49] COMPAS predicts recidivism to support a judge's parole decision. It turned out to be equally accurate for people of different ethnic background. But the mistakes were not equal in a different respect. The probability of being considered at high recidivism risk but not, in fact, reoffending was twice as high for the people of color. The opposite trend was detectable for white persons, for whom it was twice as likely that their recidivism was predicted as low risk while they, in fact, did reoffend.[50] Given this result, there is a clear need for justifying which fairness measure is used in a given algorithmic procedure. Mere technical solutions are, according to Wong, insufficient for this task. Wong draws an analogy between the task to decide about measures of algorithmic fairness and Daniels and Sabin's diagnosis, that we need to decide which reasons to accept in just deliberation. Consequently, he suggests amending the search for technical measures with political processes deciding about the criteria to be implemented in algorithmic tools.

What is particularly interesting in Wong's suggestion, however, is that he does see that "it is equally difficult for the public to know how an algorithm will affect them and … whom it will affect."[51] This clearly is the same observation that has been made for other decision procedures above, and which also puts pressure on the relevance condition in *accountability for reasonableness*. Wong makes a convincing suggestion on how to react to this opacity, namely "that consequences of an algorithm and to which groups the algorithm will affect ought to be made plain to the public in a non-technical language, especially because different fairness measures will have different implications to different groups."[52] He takes this to be an amendment to the publicity condition of the *accountability for reasonableness* approach. But before the knowledge about the effects can be made understandable, it first needs to be gathered. Thus, the requirement to measure the effect of a certain decision procedure or of specific criteria should be added to the relevance condition as it stands today. While Wong is content to follow Badano in modifying the relevance condition into a full acceptability condition, he does provide technical advice for what is an even stronger modification, namely algorithmic methods to "visualize the distributional implications of different fairness measures."[53]

This is an alternative way of seeing the role of algorithmic tools in procedural justice. They are neither a mere problem that can be corrected by the implementation of procedural justice, nor can procedural justice simply be implemented in algorithmic tools. Rather, algorithmic tools are suited to measure the effects of different procedures and thereby inform the process of choosing which reasons and principles to include in decision-making. Algorithmic tools are neither just nail nor just hammer; they are at least as much measuring tapes.

In their critique discussed above, Zimmermann and Lee-Stronach point to the epistemic insufficiency of algorithmic decision-making for questions of justice. "If we accept the claim that moral norms require justified belief or even knowledge that one is not doing wrong, then we must also accept the claim that human decision-makers should not rely uncritically on algorithmic systems when there are risks of compounding structural injustices."[54] However, the same argument can be turned upside down. Given that algorithmic tools can provide additional information about the effects of structural injustice—or, more generally, of social structures and policies—in a given decision, it would be negligent not to make use of this information.

This claim can be demonstrated regarding the real examples introduced above. The discussion of Obermeyers' study rightly points out that the reliance on biased historical data in machine-learning systems reinforces historical injustice in healthcare. It also points out that the inclusion of nonmedical data results in injustice as access to healthcare provisions is concerned.[55] There is, however, an aspect to the whole study that tends to be overlooked, namely that historical injustices would probably not have been identified and surely not quantified if the machine-learning system in question had not been employed and then brought under scrutiny. Employing the system analyzed by Obermeyer and colleagues for decision-making was questionable from the perspective of justice. However, employing it as a measure of past differential treatment, and thus as an indicator of requirements of justice, would not have been. Quite the opposite, because on this basis, Obermeyer was able to evaluate and correct the principles and criteria considered in the decision process in question.

The same could have been true for NarxCare had it been used for descriptive purposes. NarxCare and other PDMPs are ethically riskier not only because they tend to conflate law enforcement and medical aims but also because they gather information that raises data protection and privacy issues. However, PDMPs are suited to identify populations that need to invest more efforts and more risk in order to obtain medication. As such, it is widely of least benefit as a support tool for clinical decision-making; given its current mode of employment, it probably even is detrimental as a public health tool to regulate access to medication, but if used differently—which might require modification of the system itself—it could be useful to identify population who face higher hurdles in accessing healthcare services. NarxCare, according to Jennifer Oliva, identifies and bases its prediction of drug abuse risk on the following factors: "(1) the number of a patient's prescribers and dispensers, (2) the method by which the patient pays for their prescription drugs, (3) the distance a patient travels from their home for treatment and medication, and (4) the patient's criminal and sexual trauma history."[56] The problem arises when these are uncritically taken as proxies for the risk of prescription drug abuse.

It would have been possible to run the algorithm and check who is most affected by decisions based on these alleged proxies. According to the results generated by Oliva and Angela Kilby,[57] it turned out that women, people of color, people living in rural areas, the socioeconomic disadvantaged and, thus, underinsured are affected significantly more than others. Controlling for diagnosis and comorbidities would have enabled to verify that this is a bias unjustified by medical criteria. The criteria above thus could have been tested for their suitability in just decision procedures. NarxCare is not presently used for this purpose, but it most probably could be.

This is not to claim that the tools in question are tailored or ideally suited to measure bias and differential treatment. They are designed for a different purpose and, thus, will have shortcomings as epistemic tools. Neither do I claim here that identifying populations with higher hurdles in their access to, and utilization of, healthcare services *eo ipso* is sufficient for making out what justice in healthcare requires. While I do share the intuition that justice requires additional support for people who face such hurdles, this is an extra, substantial claim that is not covered by the present argument. It merely claims that algorithmic tools such as PDMPs can detect real effects that—depending on the criteria of justice one employs—might give rise to claims of justice. As will be discussed in the next section, one could do much better developing algorithms dedicated to the detection of potential unfairness. In the meantime, however, there is no reason not to use existing algorithmic tools for the epistemic purpose at hand until better options are available.

Machine-learning algorithms, when trained on the data resulting from procedurally unjust decision procedures, can—as seen above—reinforce historical injustice if employed as decision-making tools. Some authors have put it bluntly: "Much of our historical healthcare data include inherent biases from decades of a discriminatory healthcare system. […] This disparity becomes embedded in the data and therefore a model learning from these data can only regurgitate the biases in the data itself."[58] But, as the same authors demonstrate, they can do something in addition; they can make the effects of past and present social structures quantifiable if employed as measuring tools. While, historically, procedures tend to be opaque in a strong sense that differential treatment often is invisible even to the individuals making the decisions, machine-learning algorithms and their results are suited, if not sufficiently used, to breaking up this opaqueness.

AI Tools as a Measuring Tape of Justice

The use of machine-learning systems as epistemic tools would also be much more in line with carefully controlled approaches in science. Contrary to early sensationalist reports, AI is in most cases not used to decide scientific questions; it has not eliminated or replaced theory and causal hypothesis from science. Rather, AI tools are predominantly used to generate hypotheses for further analysis. Stefano Canali[59] has, for example, demonstrated how new, data-based exploratory methods complement the research process. But as exemplified in a large biomedical project, identifying relations and correlations within a data set is not the end of data-oriented research. Rather, such correlations and patterns within a data set serve to develop hypotheses and causal theories, and to test them further.

Machine-learning tools used in the context of allocational decisions in healthcare should, first and foremost, be used as epistemic tools as well. Some of the methods recently developed in the debate about algorithmic fairness seem to be ideally suited to capture specific dimensions of justice. An interesting example developed in the fairly young tradition of causal statistical modeling[60] is algorithmic recourse. In a version for causal models, algorithmic recourse identifies the means of recourse for individuals, that is, what they can do at what cost in order to repeal an algorithmic score. Julius von Kügelgen and colleagues derive two measures of fairness from this idea: one group level and one individual level. On the group level, they take an algorithm to be fair if the difference in cost for recourse for all affected individuals is zero. On the individual level, they take an algorithmic decision to be fair "if the cost of recourse would have been the same had the individual belonged to a different protected group."[61] By using a causal model of recourse, they can indeed make their predictions about the cost of recourse for different persons testable by observation. It clearly is not a result of von Kügelgen's algorithm that equal

cost of recourse is fair. The algorithm can only identify these costs. That the cost distribution is relevant for evaluations of justice is an extra, substantial claim. However, algorithmic recourse seems suited to model such a fairness claim, namely the revisions and repeal condition in Daniels and Sabin's accountability for reasonableness approach, with the slight difference that it tries to quantify the cost of revising a decision.

The study by Eliane Röösli and colleagues,[62] mentioned above, has shown how external model validation demonstrates how machine-learning systems that have been trained on the results of past decision procedures can be analyzed for differential treatments of groups with the so-called *fairness and generalizability assessment framework*.[63] They investigate a benchmarking model trained on one of the most important databases for health data, the Medical Information Mart for Intensive Care (MIMIC). By means of internal, external, and retrained model validation, they identify different representations of socioeconomic class in the model and infer "that model fairness is not guaranteed for certain ethnic and socioeconomic minority groups" and that there were "differences in patient comorbidity burden for identical model risk predictions across socioeconomic groups."[64] Thus, training on the results of past decisions and validation is suited to identify the effects of past different representations. In addition it might allow to point out limits to fairness in these past decisions and its guiding principles.

The What-If and 360 AI Fairness tools discussed by Wong,[65,66] the analysis of algorithmic recourse as demonstrated by von Kügelgen, Tina Hernandez-Boussard's MINIMAR framework, and similar tools developed by the TRIPOD-ML initiative[67] are suited to be more than mere corrections to problems of algorithmic fairness. These and similar tools can and should amend our approach to procedural justice in healthcare (and possibly beyond that).[68] They should be used to inform the decision about which principles and criteria to employ in each decision problem by making the effects of these criteria on different people transparent. Reasonable deliberation about which principles and criteria to consider in a given decision presupposes this kind of transparence. Otherwise, it is hard to see how it can be claimed that these reasons would be accepted by those affected by the decision.

It has to be admitted, however, that there is one serious limitation to the use of AI tools to measure differential effects on different people. Measuring tools themselves are not immune to the same distortions that have been observed for distributional principles and policies. What we take to be worth measuring has also developed under circumstances of unequal representation of different populations. Applying a measure developed without their participation and representation to some group is not only politically problematic; it has also recently become known to be a major scientific and moral issue in the behavioral sciences.[69] While this additional problem cannot be tackled here, a solution seems, at first hand, to lie beyond the reach of algorithmic tools and clearly in the field of participatory policies in the design of such measuring tools.

Nevertheless, AI tools seem to be a part of the problem, the nail sticking out—they do raise issues of algorithmic justice, but some of these are, first and foremost, a continuation of biases present in the already-unjust healthcare system. At the same time, they are a part of a solution, too. They cannot inform us about which principles and criteria are just in a given decision problem, but they can inform us about the effects certain principles and criteria would create. It can still be doubted that the combination of an updated procedural approach and better information suffices to solve the question of justice in healthcare. One can still insist that procedures and information alone will not provide us with substantial claims about justice.[70,71] But, at least, the amendment of the procedural approach with measuring procedures should partially cure its blindness to preexisting bias and similar effects, as the amendment of algorithmic justice was suggested to cure its lack of sensitivity to ethical deliberation.[72]

Competing Interest. The author declares none.

Notes

1. Díaz V, Viceconti M, Stroetmann K, Kalra D. *Roadmap for the Digital Patient –DISCIPULUS.* Bonn: Empirica; 2013.

2. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine* 2019;**25**(1):44–56. There is a much-repeated justification for investing heavily into AI systems in healthcare allegedly stemming from Topol's article. Here is an example for the justification: "Healthcare systems across the globe are struggling with increasing costs and worsening outcomes." (Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. *Social Science & Medicine* 2020;260:113172). But this is an inadequate reaction to what Topol can really show. Topol explicitly claims that these trends (increasing costs and worsening outcomes) are specific for the United States healthcare system. Here is the original line: "The first is a failed business model, with increasing expenditures and jobs allocated to healthcare, but with deteriorating key outcomes, including reduced life expectancy and high infant, childhood, and maternal mortality in the United States." Topol refers to two sources in order to substantiate his claim, one of them being a comparison between child mortality in the US compared to 19 other OECD countries, the other being an article with the telling title "Link between health spending and life expectancy: US is an outlier". He does go on to claim that this issue is not limited to the U.S., but he does not provide any evidence for any other national health system.
3. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine* 2020;**3**(1):43. While Chancellor and De Choudhury identify several methodological shortcomings in present studies detecting mental health status from social media postings, they also point out directions to compensate for these shortcomings and thus generate diagnostically valuable tools.
4. Heinrichs B, Eickhoff SB. Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping* 2020;**41**(6):1435–44. doi:10.1002/hbm.24886.
5. Murdoch B. Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Medical Ethics* 2021;**22**(1):122.
6. Vogt H, Hofmann B, Getz L. The new holism: P4 systems medicine and the medicalization of health and life itself. *Medicine Health Care & Philosophy* 2016;**19**(2):307–23.
7. Laacke S, Mueller R, Schomerus G, Salloch S. Artificial intelligence, social media and depression. A new concept of health-related digital autonomy. *The American Journal of Bioethics* 2021;**21**(7):4–20.
8. Beauchamp TL, Childress JF. Principles of biomedical ethics. 8th ed. Oxford, New York: Oxford University Press; 2019.
9. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics* 2019;**21**(2):E167–79.
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):447–53.
11. Oliva J. Dosing discrimination: Regulating PDMP risk scores. *California Law Review* 2022;**110**(1):47–115.
12. Kilby AE. Algorithmic fairness in predicting opioid use disorder using machine learning. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery; 2021:272.
13. Röösli E, Bozkurt S, Hernandez-Boussard T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. *Scientific Data* 2022;**9**(1):24.
14. See note 11,
15. See note 10, Obermeyer et al. 2019, at 447–53.
16. See Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics* 2021;**47**(12):e3.
17. Cowden Hindash AH, Lujan C, Howard M, O'Donovan A, Richards A, Neylan TC, et al. Gender differences in threat biases: Trauma type matters in posttraumatic stress disorder. *Journal of Traumatic Stress* 2019;**32**(5):701–11.
18. Gutmann A, Thompson D. Deliberating about bioethics. *Hastings Center Report* 1997;**27**(3):38–41.
19. Fleck LM. *Just Caring: Health Care Rationing and Democratic Deliberation*. Oxford, New York: Oxford University Press; 2009.

20. Daniels N, Sabin JE. *Setting limits fairly. Can we learn to share medical resources?* Oxford: Oxford University Press; 2002. In prior work, Daniels realized the transfer of Rawls' conception of fair equality of opportunity into the healthcare field. Accordingly, justice in healthcare does not aim at equality of opportunity *sans phrase.* Individual opportunity depends on several factors, many of which clearly lie outside the healthcare field (e.g., education), and others that might even allow for inequality (e.g., talent). Rather, justice in healthcare aims at eliminating specific undeserved restrictions on persons' opportunities and chances, namely those that result from disease and disability (Daniels N. *Just Health Care*. Cambridge: Cambridge University Press; 1985). This aim of reducing health-related restrictions to opportunity applies equally across all members of a society.
21. Badano G. If You're a Rawlsian, How Come You're So Close to Utilitarianism and Intuitionism? A Critique of Daniels's Accountability for Reasonableness. *Health Care Analysis* 2018;**26**(1):1–16.
22. See note 20, Daniels, Sabin 2002.
23. See note 20, Daniels, Sabin 2002, at 45.
24. Friedman A. Beyond accountability for reasonableness. *Bioethics* 2008;**22**(2):101–12.
25. Lauridsen S, Lippert-Rasmussen K. Legitimate allocation of public healthcare: Beyond accountability for reasonableness. *Public Health Ethics* 2009;**2**(1):59–69.
26. Ford A. Accountability for reasonableness: The relevance, or not, of exceptionality in resource allocation. *Medicine, Health Care & Philosophy* 2015;**18**(2):217–27.
27. Rid A. Justice and procedure: How does "accountability for reasonableness" result in fair limit-setting decisions? *Journal of Medical Ethics* 2009;**35**(1):12–6.
28. Ashcroft R. Fair process and the redundancy of bioethics: A polemic. *Public Health Ethics* 2008;**1**(1):3–9.
29. Hasman A, Holm S. Accountability for reasonableness: Opening the black box of process. *Health Care Analysis* 2005;**13**(4):261–73.
30. Landwehr C. Procedural justice and democratic institutional design in health-care priority-setting. *Contemporary Political Theory* 2013;**12**(4):296–317.
31. Rawls J. *A Theory of Justice*. Cambridge, MA: Belknap Press; 1971.
32. Scanlon T. *What we owe to each other*. Cambridge, MA: Belknap Press; 1998.
33. See note 20, Daniels, Sabin 2002.
34. See note 24, Friedman 2008, at 104.
35. See note 24, Friedman 2008, at 105.
36. See note 24, Friedman 2008, at 105.
37. See note 24, Friedman 2008, at 107.
38. See note 21, Badano 2018, at 1–16.
39. See note 21, Badano 2018, at 11.
40. See note 20, Daniels, Sabin 2002, at 45, emphasis added.
41. Friedman B, Nissenbaum H. Bias in computer systems. *ACM Transactions of Information Systems* 1996;**14**(3):330–47.
42. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: A critical review of fair machine learning. 2018. https://doi.org/10.48550/arXiv.1808.00023.
43. Fazelpour S, Danks D. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass* 2021;**16**(8):e12760.
44. Beauchamp TL, Childress JF. Principles of biomedical ethics. 5th ed. Oxford, New York: Oxford University Press; 2001.
45. See note 8, Beauchamp, Childress 2019, at 286.
46. Zimmermann A, Lee-Stronach C. Proceed with caution. *Canadian Journal of Philosophy* 2021;**52**(1):1–20.
47. See note 46, Zimmermann, Lee-Stronach 2021, at 2.
48. Wong P-H. Democratizing algorithmic fairness. *Philosophy & Technology* 2020;**33**(2):225–44.
49. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. 2016. https://doi.org/10.48550/arXiv.1609.05807.

50. Christian B. *The Alignment Problem: Machine Learning and Human Values*. New York: Norton; 2020.
51. See note 48, Wong 2020, at 325.
52. See note 48, Wong 2020, at 325.
53. See note 48, Wong 2020, at 325.
54. See note 46, Zimmermann, Lee-Stronach 2021, at 15.
55. Benjamin R. Assessing risk, automating racism. *Science* 2019;**366**(6464):421–22.
56. See note 11, Oliva 2022, at 200.
57. See note 12, Kilby 2021.
58. See note 13, Röösli et al. 2022, at 9.
59. Canali S. Big data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data & Society* 2016;3(2):2053951716669530.
60. Pearl J. *Causality*. 2nd ed. Cambridge: Cambridge University Press; 2009.
61. von Kügelgen J, Karimi A-H, Bhatt U, Valera I, Weller A, Schölkopf B. On the fairness of causal algorithmic recourse. 2020. https://doi.org/10.48550/arXiv.2010.06529.
62. See note 13, Röösli et al. 2022, at 24.
63. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association* 2020;**27**(12):2011–5.
64. See note 13, Röösli et al. 2022, at 24.
65. The What-If Tool; available at https://pair-code.github.io/what-if-tool/. [Last accessed 04/10/2023]
66. The AI Fairness 360 tool; available at https://aif360.mybluemix.net/. [Last accessed 04/10/2023]
67. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *The Lancet* 2019;**393** (10181):1577–9.
68. Using machine learning algorithms as epistemic tools to detect bias and injustice has probably been pioneered by Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 2018;**115**(16): E3635–44. doi:10.1073/pnas.1720347115. Garg and colleagues can identify bias in word associations using machine learning tools. A similar use for the detection of discrimination has been suggested by Heinrichs B. Discrimination in the age of artificial intelligence. *Ai & Society* 2021;37(1):143–54. doi: 10.1007/s00146-021-01192-2.
69. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? *The Behavioral and Brain Sciences* 2010;**33**(2–3):61–83; discussion: 83–135. doi:10.1017/s0140525x0999152x.
70. See note 24, Friedman 2008, at 101–12.
71. See note 28, Ashcroft 2008, at 3–9.
72. See note 48, Wong 2020, at 225–44.

Cite this article: Heinrichs J.-H (2024). Hammer or Measuring Tape? Artificial Intelligence and Justice in Healthcare *Cambridge Quarterly of Healthcare Ethics* 33: 311–322, doi:10.1017/S0963180123000257