On Algorithmic Fairness in Medical Practice

Cambridge Quarterly of Healthcare Ethics

https://doi.org/10.1017/S0963180121000839

Authors:

Dr. Thomas Grote

Email**:** thomas.grote@uni-tuebingen.de

Ethics and Philosophy Lab; Cluster of Excellence: Machine Learning: New Perspectives for Science; University of Tübingen Maria von Linden Str. 6, D-72076 Tübingen, Germany

International Center for Ethics in the Sciences and Humanities (IZEW); University of Tübingen; Wilhelmstraße 19, D-72074 Tübingen, Germany

Dr. Geoff Keeling

Email: gkeeling@stanford.edu

McCoy Family Center for Ethics in Society, Stanford University, 559 Abbot Way, Stanford CA 94305, United States

Institute for Human-Centered Artificial Intelligence, Stanford University, Cordura Hall, 210 Panama Street, Stanford, CA 94305, United States

Abstract: *The application of machine learning technologies to medical practice promises to enhance the capabilities of healthcare professionals in the assessment, diagnosis, and treatment, of medical conditions. However, there is growing concern that algorithmic bias may perpetuate or exacerbate existing health inequalities. Hence it matters that we make precise the different respects in which algorithmic bias can arise in medicine, and also make clear the normative relevance of these different kinds of algorithmic bias for broader questions about justice and fairness in healthcare. In this paper, we provide the building blocks for an account of algorithmic bias and its normative relevance in medicine.*

Keywords: *Fairness; Machine Learning; Algorithmic Bias; Discrimination; Medical Practice*

1. Introduction

Recent developments in machine learning have seen the application of data-driven technologies to the assessment, diagnosis, and treatment, of medical conditions. The potential benefits of these technologies are manifold. Consider some examples. In radiology, dermatology and ophthalmology, machine learning algorithms have achieved a level of detective accuracy comparable to that of medical professionals in the classification of medical conditions.[1] For instance, an algorithm trained by DeepMind was able to predict the risk of an eye developing 'wet' age-related macular degeneration within the next six months at a higher level of accuracy than most experts.[2] In a similar manner, said algorithms can be utilised for the prediction of mental health problems. Notably, these algorithms can be also used to predict the occurrence of mental-health problem, such as predicting a post-traumatic stress course for patients suffering from a traumatic event.[3]

These algorithms thus lay the ground for advanced decision support systems that enable physicians to minimise diagnostic errors. Finally, there are high hopes that machine learning will pave the way for prescriptive analytics, whereby we mean recommender systems which can provide personalised treatment plans supplemented with relevant medical evidence to assist doctors and patients in making informed treatment decisions.[4]

As the predictive accuracy of diagnostic systems surpasses that of medical professionals, and decision support algorithms facilitate increasingly personalised healthcare, it is only a matter of time before machine learning systems become an integral part of medical practice. While chances are high that the incorporation of machine learning will improve the accuracy of medical decisions, there is also a rich body of evidence indicating that machine learning algorithms are at risk of being unreliable outside of training conditions.[5] This is particularly evident for biases against social minority groups, such as women or people of non-European ancestry.[6] The bottom line is that due to these biases they are at risk of receiving unfair treatment in clinical settings.

The implementation of machine learning thus threatens to exacerbate existing inequalities in healthcare, prompting urgent reflection on the means to ensure algorithmic fairness in medical practice. That said, whereas there has been a surge of interest in the (philosophical) ethics of medical AI,[7] remarkably little has been written on the issue of algorithmic fairness in in medical practice. Why is that? Our conjecture is that the oversight can at least in part be attributed to the predominantly individualist approach in medical ethics, that makes it difficult to account for issues of healthcare inequalities and distributive justice. As a consequence, much of the current literature is focussed on examining the potential threats of machine learning for patient autonomy. By contrast, we address the problem of algorithmic fairness at the population level while making clear why our considerations matter for medical practice.

With that in mind, our paper has three objectives:

(i) We try to provide an account of the ways that algorithmic discrimination manifests in medical practice.

(ii) We try to examine the underlying mechanisms of algorithmic bias.

(iii) We try to identify the appropriate normative standards for fair algorithmic decision-making in medical practice.

In addition, one upshot of the paper is that it ties together certain themes in the emerging literature on 'fairness in machine learning' with the philosophical debate on justice in healthcare. The remainder of the paper will be structured as follows: In section II, we give an outline on the different mechanisms of algorithmic bias in medical practice, whereby we distinguish between *formal*, *substantive*, and *normative* notions of algorithmic bias. Thereafter, section III will discuss to what extent different standards of fairness help to counteract the threat of algorithmic discrimination. Here, we argue that fairness cannot be restored merely by mitigating the differences in the algorithm`s predictive accuracy for different demographics. Instead, a wider array of normative criteria needs to be taken into account.

By discussing the constraints of formal accounts of algorithmic fairness, we hope to provide building blocks for a comprehensive theory of algorithmic fairness in medical practice.

2. Mechanisms of Algorithmic Bias in Medical Practice

Machine learning algorithms are statistical models that are trained to perform domain-specific tasks such as prediction and classification. For example, a machine learning algorithm in medicine might be trained to answer questions such as 'Is this macule a tumour?' or 'Is this patient at risk of having a stroke?'. What sets machine learning algorithms apart from handcrafted statistical models is that developers will not manually set the model parameters. Instead, in a supervised learning setting, the algorithm learns through a process of trial-and-error training which modelling parameters are best suited to the task at hand.[8] The upshot is that these models prove to be more flexible than handcrafted models, at least under the assumption that the target category to predict is narrowly defined and that large volumes of training data are available.

Consider an example. Suppose that a hospital wants to train an algorithm to predict the risk of patients suffering a heart attack within the next year. The developer will first assemble a large dataset that contains information about various features of patients and their medical histories. For example, their gender, ancestry, medical conditions, weight, dietary habits, family status, employment, and so on. The data will also include information about whether each patient suffered a heart attack within a year of the data being recorded. The developer will then build an algorithm that takes as its inputs the data of particular patients, and outputs a prediction about whether the patient will suffer a heart attack within the next year. The parameters in the model will initially be set randomly. The developer will then run a trial-and-error training process to fine-tune the parameters of the model. Each time the algorithm correctly identifies whether a patient in the training data had a heart attack within a year, the parameters are left alone; and each

time the algorithm outputs the wrong prediction, its parameters are revised to compensate for the error. Over time, the algorithm will converge on a set of parameters that reliably predicts the risks of patients of having a heart attack within the next year conditional on a patient's input data. The art and craft of developing machine learning algorithms amounts to training a model, which captures the regularities in the training data, while also generalizing well for data that it has not encountered in training (to prevent overfitting).

Machine learning algorithms lend themselves naturally to two applications in medicine: medical diagnosis via image recognition, and the prediction of health risks. In both areas, there is a wealth of studies in which the relevant algorithms have exceeded the predictive and diagnostic capabilities of medical experts.[9] However, these models are vulnerable to biases when assessing patient-groups that are insufficiently well-represented in the training data. In that case, the algorithm will perform worse when predicting health-related properties for a given underrepresented group of patients. Oftentimes, these biases will affect members of salient social groups, meaning groups who are vulnerable across a wide range of social contexts, such as women or people of a certain race or ethnicity.[10] In the context of medical practice, algorithmic bias has therefore an inherent moral dimension.

Before we discuss the moral significance of algorithmic bias, it is necessary to make some conceptual clarifications. For a start, we distinguish between three kinds of algorithmic bias: formal, substantive, and normative. In drawing this distinction, it is helpful to consider two examples from medical practice which are fictional, yet are likely to occur in spite of the functioning of current machine learning algorithms:

Medical Diagnosis: A hospital utilises an algorithm to examine potential cases of skin-cancer based on medical images. As it turns out, its general accuracy is 20% lower when assessing images from Black patients as compared to White patients.

Prediction of Health Risks: A hospital decides to use an algorithm to monitor the health risks of patients. Once a certain threshold has been reached, a given patient will be referred to intensive care. When revalidating the algorithm, it becomes apparent, that in order to reach the very same risk score, women are, on average, in a more health-critical state than men before being referred to intensive care.

The formal notion of algorithmic bias holds that an algorithm is biased if, and only if, there exists some property P such the algorithm's predictive accuracy is worse for patients with property P than for patients who lack property P. The subpar performance of an algorithm for a certain group of patients can be attributed to some form of model bias. What this means is that when examining some ambiguous input, a given model will favour certain interpretations over others or exclude certain interpretations altogether.[11] Consequently, the predictive accuracy will be lower for patients with property P, as compared to other patients. In machine learning parlance, we might also say that the model underfits for said group of patients. Notice that this kind of algorithmic bias is purely extensional, in the sense that if the algorithm is biased with respect to property P, then it will also be biased for any property Q which is coextensive with P. Thus, formal algorithmic bias may be uninteresting from the point of view of justice and fairness, insofar as certain properties for which the algorithm is biased may have no independent moral significance.

By contrast, the substantive notion of algorithmic bias is hyperintensional. That is, an algorithm might be biased with respect to property P, but not biased with respect to Q, even though P and Q are necessarily coextensive. To make this clear, let us take a step back and consider, why algorithmic bias arises in the first place. Here, we can distinguish between three different sources:[12]

(i) *Population Inequity*: patients with the property P, might be underrepresented in the training data, which is why the model underfits for them.

(ii) *Human Bias*: patients with the property P, might have been treated worse than other groups by medical professionals and these tendencies carry over to the training data.

(iii) *Problem Formulation*: the developers might have defined the algorithm`s target category in a way that patients with the property P, are disadvantaged in comparison to other groups.

Now, in *Medical Diagnosis*, what explains why the algorithm is biased against Black patients is that the training set was insufficiently diverse with respect to skin colour. Thus, the property of having black skin can feature in explanations of why the algorithm is less accurate for Black patients (whereas any other property that is necessarily coextensive with having black skin would not do this). Due to a lack of exposure to medical images from patients with black skin-color, the model overfits for some populations (i.e. Whites) while underfitting for Black patients. Conversely, there may exist properties that are necessarily coextensive with being Black that do not explain the algorithm's bias.

It needs to be emphasized that *Population Inequity* in datasets poses a huge problem for medical machine learning, even more so in the case of the *Prediction of Health Risks*. As health risks can be attributed to a combination of different factors, e.g. gender, ancestry, lifestyle choices, dietary habits or the environment, it is a necessary prerequisite to have large quantities of data from diverse populations in order for an algorithm to predict equally well for different demographics.[13]

The problem, however, is that depending on where the medical data has been collected, some populations will almost inevitably be underrepresented. As an example, an academic hospital somewhere in Central Europe will most likely have many fewer Black patients than Whites. This leaves the developers of the algorithm with two options. First, they might try to merge medical data from different sources (i.e. other academic hospitals around the globe). Second, they over-proportionately collect medical data from underrepresented populations (i.e. by

persuading them to donate their data). The crux is that both options face moral and legal constraints.

For a start, it is not adequately understood who owns (or should own) said medical data.[14] The more concerning issue however is that medical data contains sensitive information on patients. Although medical data are stored anonymously, there is an underlying worry that machine learning methods can be used to re-identify patients. A worst-case scenario might be that the genetic data of individual leaks to the public and an insurance company or an employer then find out of an applicant's pre-dispositions to diseases. By collecting and sharing medical data, patients are therefore exposed to risks concerning their privacy rights.[15] In the European context, where the sovereignty of citizens over their personal data might be even deemed as a foundational right, there are thus strong barriers to collecting and sharing medical data.[16]

Especially for social minority groups, this creates a dilemma: if there is not enough medical data available, there is a high chance that the algorithm will generate more false findings for them. At the same time, by trying to over-proportionately collect their medical data to compensate for *Population Inequity*, the privacy rights of individual members of given social minority groups are disproportionately exposed to risks. Either way, they are at risk of being treated unfairly. As of now, many of the currently existing medical datasets only include information on a rather homogeneous population.[17]

Concerning *Human Bias*, there are two sorts of problems. First, it is well-established that medical judgement is affected by racial and gender bias. For instance, Black patients tend to be underdiagnosed and undertreated for pain in comparison to White patients.[18] Second, in clinical trials ethnic minority groups and especially women tend to be underrecruited and gender-related data is underreported. We thus have little information about the effects of medical drugs on women.[19] And of course, the same argument holds true for children! When

training an algorithm which assesses the health risks of patients, this disparity will reflect in the dataset.

Of all the sources of bias which we discuss, *Problem Formulation* might be the most intricate one. The issue may be subsumed as follows: To predict a health-related property P, the algorithm makes inferences from a set of proxy features Q, which act as predictors. The selection of these proxy features can give rise to biases in that a given prediction will be more accurate for some demographics than for others. This has become apparent due to a widely acclaimed study by Obermeyer et al.,[20] in which the researchers scrutinized an algorithm that is widely used in the United States to predict health care needs of patients. As evidence shows, the relevant algorithm was miscalibrated for Black patients. They were, at a given risk score, considerably sicker than White patients. One of the key-factors which the researchers attributed to this miscalibration has been that the health-costs of individual patients were used as the main predictor. Due to various socio-economic factors, Black populations have less access to healthcare than Whites. In consequence, they might visit the hospital less frequently than their White counterparts and this leads to disparities in health-costs and the perception that they are less interested in their own health. This constitutes a form of indirect discrimination.[21]

There are two lessons to be drawn from the study. The first one is simple: developers should be wary of the larger social context when choosing the target category for an algorithm. By contrast, the second one is that there may be profound problems in the way that machine learning algorithms operate, namely by making predictions in virtue of exploiting correlations in the data while being unable to capture potentially discriminatory factors. We will address this issue in the subsequent chapter. In sum, the substantive notion helps us to understand the various pragmatic and normative constraints which give rise to algorithmic bias. It does not, however, grasp the whole morally normative dialectic of algorithmic bias.

This makes it necessary to introduce the normative notion of algorithmic bias. Thus, the objective is to understand what it is that makes (certain types) of algorithmic bias morally reprehensible in the context of medical practice. As a starting point, there is a close link between bias and unfair treatment as is showcased by the example of *Medical Diagnosis*, where Black patients receive worse medical treatment due to algorithmic bias. However, not every unfair treatment against a given certain social group might be morally reprehensible. Suppose that instead of discriminating against women or Black patients, the algorithm`s accuracy would be worse when examining data from golf players as compared to non-golf players. The bias might be deemed to be a flaw in the model, but not necessarily as a moral wrong. The reason is that being a golf-player is not a morally relevant property in the sense that ethnicity, race, or gender are.[22]

What should be the morally relevant reference class within the context of algorithmic fairness in medical practice? In the literature on justice in healthcare, we basically find two answers: either the relevant inequalities are defined in terms of a difference between two or more social groups[23] or, since health is not a property of groups, it is assumed that health differences across individuals are what is what matters morally.[24] Both options have different implications with regards to the metric of fairness. We might either compare the performance of the algorithm for members of social groups or we might measure how individuals with similar properties are being treated (this is typically being referred to as Individual Fairness). In either case, some higher-order normative problems arise. For instance, how should we sub-divide a given population into sub-groups[25] and which properties of individuals should be determinative when measuring the performance of the algorithm?

3. From Algorithmic Bias to Fairness

In this section, we now step fully onto normative terrain as our objective is to develop an account of fairness for algorithmic decision-making in medical practice. We start with some brief remarks with regards to fairness in machine learning.

As machine learning has become ubiquitous in many societal domains, consequential decisions are increasingly made by algorithms. Oftentimes, the rationale is that the involvement of machine learning will foster fairer decisions. Suffice it to say, various reports on algorithmic bias against salient social groups in hiring, policing or judicial decisions cast doubts on whether this optimism is still apt. When considering the detrimental effects of machine learning on social justice, a study on the COMPAS-algorithm by ProPublica proves to be a particularly instructive example.[26] The algorithm, which is used in many states in the US, is utilized to assess the risks of criminals to become recidivist and thus it informs sentencing decisions in court. As the study revealed, the algorithm reflects racial bias in that the algorithm overpredicted the risks of Black defendants (the ratio of false positives was twice as high as compared White defendants) while underpredicting the risks for White defendants (the ratio of false negatives was twice as high as compared to Black defendants).

It might be fair to say that example of the COMPAS-algorithm has become an anchor point in the debate on fairness in machine learning.[27] For the aims of our paper, it proves to be an interesting contrast foil when comparing the requirements of algorithmic fairness in the contexts of judicial- and medical decision-making.

Fairness in Machine Learning has emerged as an interdisciplinary field of research that is driven by two related questions:

(i) Which standard of algorithmic fairness should be adopted within a given context?

(i) How can the given standard of fairness be implemented into algorithmic decision-making?

The first question is concerned with the normative side, whereas the second`s emphasis lies on technical feasibility. As an example, how can different notions of fairness such as 'equality of opportunity' or 'equality of outcome' be formalized? Our concern lies in the normative claims of fair machine learning. That said, the aim is not to give a comprehensive overview of the debate, but we will focus on those aspects which we consider pivotal for algorithmic fairness in clinical settings.

Our starting point is that fairness is not a uniform concept. Instead, there are a bundle of different criteria about what should count as fair and these refer to distinct normative ideas – an insight which may not exactly come as a shock to philosophers.

a) *Distributive Accounts of Algorithmic Fairness in Medical Practice*

The literature on fairness in machine learning is vast and various accounts on how to ensure fairness in algorithmic decision-making are still contested.[28] With that in mind, most of these accounts can be classified as distributive accounts of fairness, whereby an algorithmic decision counts as unfair if X has been treated worse than Y without good cause. Consequently, the currency of fairness is that of equality of outcome. Where there is disagreement within the literature is what should be the reference classes for the comparison. We might either compare the fairness in treatment across individuals with similar properties or we take social groups as the point of reference. Moreover, one might also consider how X has been treated by an algorithm, as compared to groups or individuals not affected by the algorithm.

Aside from the selection of the reference classes, there is also widespread disagreement with respect to how the equality of outcome should be measured. Therefore, different technically nuanced solutions have been developed, such as 'equalized odds', to 'equality of opportunity'. To illustrate the differences between the two concepts, 'equality of opportunity' might be conceived as a relaxation of 'equalized odds'. The later tries to make sure that both the true

positive rate and the false positive rate for different groups is the same, while the former focuses on making the true positive rate to be the same across groups. In plain words, equal opportunity ensures that the probability of granting a loan to individuals that would 'repay the loan' should be the same across different social groups. Equalized odds also need to ensure that the probability of making an error by denying the loan to people that would repay should be the same across groups.

In our view, the most promising candidate as a metric of fairness within the context of medical practice is what Deborah Hellman refers to as 'error ratio parity'. The underlying idea is that since medical diagnosis will not yield bulletproof findings, the harms need to be balanced for different subjects of algorithmic decision-making. Since reporting the false positive and false negative rate has become the default in medical testing, as a metric of fairness, the error ratio parity does align with the norms of medical reasoning. Notably, the error ratio parity does not assume that the ratio of false positives and false negatives is identical for different groups but that the harms are distributed fairly. If an algorithm which diagnoses skin-cancer generates a slightly higher rate of false negatives for Black patients while generating a slightly higher rate of false positives for White patients, the outcome is fair, if the harms have been distributed equally. This however raises the question how to weigh the harms of false positives and false negatives in medical practice.[29]

Notwithstanding its appeal, it also needs to be emphasized that as a standard of fairness, error ratio parity has its shortcomings. In particular, it is a purely formal account of fairness, which does neither tell us something about how to choose the reference classes, nor does it say anything about the broader aims of what fairness in medical practice ought to achieve. And to sharpen things, on a purely formal account, we could also achieve fair outcomes between two groups A and B, simply by making the position of B worse off.[30]

b) *Procedural Fairness*

A first step to develop a more substantive theory of fairness within the context of medical practice is to widen the scope by not only considering whether a given decision has been fair for the parties involved in terms of equality of outcome, but also whether the procedure is such, that a fair outcome is likely to obtain.[31] With regards to the application of machine learning for judicial- or policing-decisions, the most common route is to regulate the use of certain labels in the dataset. For instance, the discrimination law in the US prescribes that information concerning a person's racial identity should be protected.[32] It needs to be pointed out, that this approach has clear limitations as socio-economic factors are oftentimes highly correlated with a person's racial ancestry. In a city with high racial segregation, data concerning a person`s zip code or her income can be indicative of her race. Hence, there is a threat that making inferences on these proxy data can lead to racial discrimination. Inspired by a recent paper from Sandra Wachter and Brent Mittelstadt,[33] the task to ensure procedural fairness of algorithmic decision-making within healthcare consists in: (i) identifying which data form a normatively acceptable basis from which to draw inferences, (ii) making clear, why the data is both relevant and normatively acceptable for a certain type of prediction and (iii) whether the data and methods used to draw the inferences are accurate and statistically reliable.

It might be fair to say that the relevant task proves to be challenging on many levels. For a start, it is controversial to what extent racial classification is epistemically useful, if any, within the context of medicine.[34] At the centre of the debate lie the questions, whether racial categories possess biological reality and if so, whether these allow for reliable inferences when predicting health-related properties. Do they for example help us to predict the health risks for members of salient social groups more accurately with regards to certain types of diseases? Likewise, the same problems arise with regards to gender-related categories. While we do not attempt to resolve the debate in this paper, let us raise a few normative concerns about the use of racial categories with regards to the application of machine

learning in medical practice. The first concern is that by making inferences based on racial categories, we run a risk of approaching health-related differences between races in biologically essentialist terms. In that respect, the study from Obermeyer et al.,[35] which we discussed in the previous section exemplifies how health-related characteristics are oftentimes grounded in social factors. Hence, the worry is that by incorporating said racial categories in medical datasets, they might act as colliding/confounding variables, which mask structural injustices.

This also gives rise to a second (higher-order-) concern, which relates to the functioning of machine learning algorithms within the context of healthcare. Broad brush, to predict a health-related property X, the algorithm makes inferences by exploiting correlations within a given set of data. However, the medical data is not underpinned by a background theory and is stripped off the larger social context (i.e. what is the meaning of certain labels? What are potentially discriminatory relationships in the data?). Furthermore, due to the complexity of the machine learning model, the causal pathways which give rise to a given prediction oftentimes are elusive to developers or medical professionals.[36] A solution which has recently proposed by Matt Kusner and Joshua Loftus[37] is to run different types of causal tests which try to identify a machine learning model`s causal pathway. By knowing of a given algorithm`s causal pathway, inappropriate relationships can be captured and mitigated. In their words: "[I]f 'health-care access' is not observed in a model to predict 'health need', then it is crucial to identify any potential impacts it might have on 'health-costs' as well as how it is affected by 'ethnicity'"[38].

c) *Revisiting the Normative Foundations of Algorithmic Fairness in Medical Practice*

What is distinctive about fairness as a good is that it is a second-order good.[39] Theories of fairness disagree about how first-order goods such as health, utility, money, life years, and so on, ought to be distributed between persons. We have seen that theories of fairness in machine learning have sought to locate fairness and unfairness in facts about distributions

that pertain to algorithmic decisions. The two candidate distributions are the *ex post* distribution of errors, that is, the relative frequencies of false positive and false negative errors across different sub-groups in the population; and the *ex ante* chances that persons in different sub-groups will be subject to these errors. Roughly, fairness in the first distribution reflects equality of outcome; and fairness in the second distribution reflects procedural fairness. The idea is that fairness demands parity in error frequencies either before or after the decisions have been made. This approach gets a lot right. But our aim in this last section is to develop a more plausible normative framework in which to understand algorithmic fairness in medical decisions.

It is helpful to distinguish *algorithmic decisions* and *final decisions*. When there is a human in the loop, these two decisions are separate. For example, the COMPAS algorithm makes a prediction about an individual's risk of recidivism. This is the algorithmic decision. Then a judge makes the final decision about the appropriate sentence for that individual. When there is no human in the loop, the algorithmic decision is identical to the final decision. For example, a triage algorithm that allocates hospital beds to patients based on relevant features of the patient's history makes both an algorithmic decision and a final decision about which bed the patient is allocated. The first point we want to make is this: When there is a human in the loop, the fairness of the algorithmic decision is of merely instrumental importance. To illustrate: That a diagnostic algorithm is unfair in the sense that it is less accurate for patients in a particular minority group is morally uninteresting if the extent of the bias is known and the bias is properly accounted for in the eventual treatment decision. When there is no human in the loop, the algorithmic decision is the final decision, and thus it has intrinsic moral significance. Thus, what a fair algorithmic decision looks like will crucially depend on the role of the algorithm.

It is also helpful to distinguish *algorithmic bias* and *decisional fairness*. On one hand, algorithmic bias is a descriptive or empirical feature of the algorithm that might roughly be

understood as disparities in predictive accuracy over normatively salient sub-groups in the population. On the other hand, decisional fairness is an evaluative feature of the final decision made, i.e. whether or not the final decision is fair. The relation between algorithmic bias and decisional fairness depends on whether the system in question has a human in the loop or not. In both cases, the relation is an explanation relation. When there is no human in the loop, algorithmic bias can fully explain decisional unfairness. When there is a human in the loop, algorithmic bias provides a partial explanation of decisional unfairness. These distinctions put us in a better place for developing a normative framework for algorithmic fairness.

There are, in effect, three questions that need to be answered by any theory of fair algorithmic decision-making in medicine. First, what makes a decision fair or unfair in this context? Second, how exactly does algorithmic bias feature in explanations as to why a particular decision is fair or unfair (in both human in the loop cases and fully automated decisions)? Third, given the role that algorithmic bias plays in explaining decisional unfairness, what concrete steps can clinicians and algorithm designers take to ensure that decisional unfairness is counteracted? In what remains of this section, we shall say something about each of these questions.

First, what makes a decision fair? There is a minimal sense of fairness according to which a decision is fair if, and only if, the rules governing that decision are applied impartially. Brad Hooker[40] calls this *formal fairness*. In general, formal fairness is necessary but insufficient for a decision to be fair. The problem is that rules can be applied impartially where those rules are *substantively unfair*. For example, we can imagine a fully-automated triage algorithm that systematically prioritises White patients. This algorithm is formally fair if the rule is applied without making exceptions for particular non-White patients here and there. But the rule is substantively unfair because it is based on a morally irrelevant distinction. Exactly what substantive fairness consists in is a matter of dispute.[41] But the leading account is due to

John Broome.[42] According to Broome, fairness consists in the proportional satisfaction of claims, where claims are to be understood as moral duties owed to the relevant parties[43].

Second, how does algorithmic bias feature in explanations of unfairness? Questions about resource allocation in medicine are difficult.[44] But what is ultimately at issue from the point of view of fairness is that medical resources are distributed so as to give the greatest proportional satisfaction to the moral claims that members of the population have on those resources. Here a person has a claim on a medical resource if the patient has a medical *need* for that resource.[45] Given this account of fairness, it is obvious how algorithmic bias could in principle provide an explanation of why resources are not being allocated in accordance with needs. The algorithm is – directly or indirectly – assessing a patient's need for a resource, be it a treatment of a particular kind, or a drug, and so on. Thus, if an algorithm has better predictive accuracy for some normatively salient groups in the population, it follows immediately that unfairness will result in the allocation of resources. The situation is more complex when there is a human in the loop. However, it is more or less clear that the fact of algorithmic bias, in conjunction with the failure on the part of clinicians to recognise and correct for that bias in the allocation of medical resources, explains how medical resources might end up being distributed unfairly.

Third, what concrete steps can clinicians and the designers of medical algorithms take to ensure algorithmic fairness? The first is to develop plausible metrics for algorithmic bias, and to examine the ethical costs and benefits of different methods for collecting the data that is required to make these assessments. A great deal of progress has already been made in the literature on fairness in machine learning. Second, it matters that the fairness of algorithmic decisions is understood in the broader context of fair resource allocation in medicine. This is ultimately what matters from the point of view of fairness. Thus, it is important to make precise existing health inequalities and what the causes of these inequalities are. Third, regulators and the developers of predictive algorithms in medicine need to pay close

attention to the role that these algorithms will play in the process of medical decision-making. When developing medical algorithms, there is good reason to bring in stakeholders from different parts of the medical profession, including clinicians, patients, and other healthcare workers. What is required to fully understand how algorithms might contribute to or mitigate unfair resource allocation in medicine is a broad and inclusive discussion about these algorithms and their role in medical decisions. This dialogue will help to identify the means to account for algorithmic biases in the broader decisional context to ensure a fair allocation of medical resources.

What we are suggesting matters both for fairness in the allocation of medical resources; and for the related problem of structural injustices in medicine. As we increase the use of novel machine learning technologies in medical decision-making, there is scope both to rectify existing disparities in health across normatively salient groups of the population, and, conversely, to exacerbate these health inequalities. By paying close attention to fairness in the allocation of resources, and the role of algorithmic bias in explaining the fairness or unfairness of resource allocation decisions, we are well placed to reduce and perhaps eliminate these unjust health inequalities.

4. Conclusion

The incorporation of machine learning in medical practice threatens to exacerbate health inequalities across salient social groups. In this paper, we have identified different mechanisms which give rise to algorithmic bias and have discussed pragmatic and moral constraints of bias mitigation strategies. Building thereon, we tried to lay the ground for an account, which involves both formal and normative criteria for measurements of outcome and decisional procedures for algorithmic fairness within the context of medical practice. Finally, we put these criteria into a wider context by reflecting on the moral aims of fairness

and by pointing out possible directions which developers of algorithms and other stakeholders might pursue to remedy potentially discriminatory effects brought about by implementing machine learning systems into healthcare settings. In that respect, we hope to have defined the parameters within which a theory of algorithmic fairness in medical practice should operate. Yet, we are happy to admit that before our account developed here can be translated into a set of actionable policies, much further work will be required, ideally in a joint effort alongside medical professionals, machine learning researchers and policymakers.

Literature:

[1] For a recent meta-analysis, see: Liu Xiaoxuan, Faes Livia, Kale Aditya U, Wagner Siegfried K, Fu Dun Jack, Bruynseels Alice, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 2019;1(6):e271-e297. doi: 10.1016/S2589-7500(19)30123-2.

[2] Yim Jason, Chopra Reena, Spitz Terry, *et al.* Predicting conversion to wet age-related macular degeneration using deep learning. *Nat Med 2020* 26**,** 892–899. https://doi.org/10.1038/s41591-020-0867-7

[3] Schultebraucks Katharina, Shalev Arieh, Michopoulos Vasiliki. *et al.* A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor. *Nat Med* 2020 26, 1084–1088. https://doi.org/10.1038/s41591-020-0951-z

[4] Zhang Sushen, Bamakan Sushen, Qu Quiang, Li Sha. Learning for Personalized Medicine: A Comprehensive Review From a Deep Learning Perspective. IEEE Reviews in Biomedical Engineering 2019;12:194–208. doi: 10.1109/RBME.2018.2864254.

[5] Geirhos Robert, Jacobsen, Jörn-Henrik, Michaelis Claudio, et al. Shortcut learning in deep neural networks. 2020 arXiv:2004.07780 **[cs.CV]**

[6] See Caliskan Aylin, Bryson Joanna J, Narayanan Arvind. Semantics derived automatically from language corpora contain human-like biases. Science 2017;356(6334):183. doi: 10.1126/science.aal4230; Obermeyer Ziad, Powers Brian, Vogeli Christine, Mullainathan Sendhil. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447. doi: 10.1126/science.aax2342.

[7] Ploug Thomas, Holm Soren. The right to refuse diagnostics and treatment planning by artificial intelligence. Medicine, Health Care and Philosophy 2020;23(1):107–14. doi: 10.1007/s11019-019-

09912-8; Grote Thomas, Berens Philipp. On the ethics of algorithmic decision-making in healthcare. Journal of Medical Ethics 2020;46(3):205. doi: 10.1136/medethics-2019-105586; McDougall Rosalind J. Computer knows best? The need for value-flexibility in medical AI. Journal of Medical Ethics 2019;45(3):156. doi: 10.1136/medethics-2018-105118; Bjerring Jens C, Busch Jacob. Artificial Intelligence and Patient-Centered Decision-Making. Philosophy & Technology 2020. doi: 10.1007/s13347-019-00391-6.

[8] Cf. Jordan Michael , Mitchell Tom. Machine learning: Trends, perspectives, and prospects. Science 2015;349(6245):255. doi: 10.1126/science.aaa8415; LeCun Jan, Bengio Yoshua, Hinton Geoffrey. Deep learning. Nature 2015;521(7553):436–44. doi: 10.1038/nature14539.

[9] See for reviews: Topol Eric. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25. doi: 10.1038/s41591-018-0300-7; see also note 3, Liu et al. 2019; for a philosophical exposition, see: Grote Thomas, Berens Philipp. Uncertainty, evidence, and the integration of machine learning into medical practice. The Journal of Medicine and Philosophy (forthcoming).

[10] Cf. Lippert-Rasmussen Kalle. Born Free and Equal?: A Philosophical Inquiry Into the Nature of Discrimination. Oxford, New York: Oxford University Press; 2014: 30-36.

[11] Sinz Fabian H, Pitkow Xaq, Reimer Jacob, Bethge Matthias, Tolias Andreas S. Engineering a Less Artificial Intelligence. Neuron 2019;103(6):971. doi: 10.1016/j.neuron.2019.08.034.

[12] Cf. Green Brian. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 594–606.

[13] Shilo Smadar, Rossman Hagai, Segal Eran. Axes of a revolution: challenges and promises of big data in healthcare. Nature Medicine 2020;26(1):29–38. doi: 10.1038/s41591-019-0727-5.

[14] Cf. Ballantyne, Andrea. How should we think about clinical data ownership?. Journal of Medical Ethics 2020 (online first)

[15] Mittelstadt Brent D, Floridi Luciano. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Science and Engineering Ethics 2016;22(2):303–41. doi: 10.1007/s11948-015-9652-2; Véliz Carissa. Not the doctor's business: Privacy, personal responsibility and data rights in medical settings. Bioethics 2020;n/a(n/a). doi: 10.1111/bioe.12711.

[16] Hummel Patrick, Braun Matthias, Dabrock Peter. Own Data? Ethical Reflections on Data Ownership. *Philos. Technol.* 2020. https://doi.org/10.1007/s13347-020-00404-9

[17] See note 13, Shilo et al. 2020.

[18] Hoffman Kelly M, Trawalter Sophie, Axt Jordan R, Oliver M Norman. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proceedings of the National Academy of Sciences 2016;113(16):4296. doi: 10.1073/pnas.1516047113.

[19] Cf. Holdcroft Anita. Gender bias in research: how does it affect evidence based medicine? Journal of the Royal Society of Medicine 2007;100(1):2–3. doi: 10.1177/014107680710000102.

[20] See note 6, Obermeyer et al. 2020.

[21] Cf. See note 10, Lippert-Rasmussen 2014:54-56.

[22] Cf. See note 10, Lippert-Rasmussen 2014:30-35.

[23] Hausman Dan M. What's Wrong with Health Inequalities?*. Journal of Political Philosophy 2007;15(1):46–66. doi: 10.1111/j.1467-9760.2007.00270.x; Marmot Michael. Social determinants of health inequalities. The Lancet 2005;365(9464):1099–104. doi: 10.1016/S0140-6736(05)71146-6.

[24] Lippert-Rasmussen, Kalle. When group measures of health should matter. In: N Eyal, S Hurst, OF Norheim, D Wikler, eds. *Inequalities in Health: Concepts, Measures, and Ethics*. New York: Oxford University Press; 2013:52-65.

[25] See note 24, Lippert-Rasmussen 2013.

[26] Angwin, Julia., Larson Jeff, Mattu Surya, Kirchner Lauren. Machine Bias; 2016 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing: last retrieved on April 1st, 2020.

[27] Barocas Solon, Selbst Andrew. Big Data's Disparate Impact. 104 California Law Review 671;2016.

[28] See for an overview: Mehrabi Ninareh, Morstatter Fred, Saxena Nripsuta, Lerman Kristina, Galstyan Aram. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635;2019; Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning: fairmlbook.org; 2019.

[29] Cf. Hellman Deborah. Measuring Algorithmic Fairness. Virginia Public Law and Legal Theory Research Paper No. 2019-39 [Internet];forthcoming. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418528.

[30] Thanks for an anonymous reviewer for pointing this out.
[31] Cf. Rawls John. A Theory of Justice. Cambridge, MA: Harvard University Press; 1971:§14.

[32] Cf. See note 27, Barocas, Selbst 2016

[33] Wachter Sandra, Mittelstadt Brent. A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI. Columbia Business Law Review 2018(2):443–93.

[34] Spencer Quayshawn NJ. A racial classification for medical genetics. Philosophical Studies 2018;175(5):1013–37. doi: 10.1007/s11098-018-1072-0; Yudell Michael, Roberts Dorothy, DeSalle Robert, Tishkoff Sarah. Taking race out of human genetics. Science 2016;351(6273):564. doi: 10.1126/science.aac4951.

[35] See note 6, Obermeyer et al. 2020.

[36] The complexity of machine learning models raises wider epistemic and ethical concerns with respect to the non-interpretability of algorithmic decisions. We will not address these issues in this paper. Sullivan, Emily. Understanding from machine learning models. British Journal for the Philosophy of Science; 2019 (online first) gives a good introduction to the epistemic side of the topic.

[37] Kusner Matt J, Loftus Joshua. The long road to fairer algorithms. Nature 2020;578(7793):34–6.

[38] See note 37, Kusner, Loftus 2020:35.
[39] Broome, John. Weighing Lives. Oxford: Oxford University Press, 2004:38.
Broome, John. Weighing Goods. Oxford: Blackwell Publishers, 1991.

[40] Hooker Brad. Fairness. Ethical Theory and Moral Practice 8, 2005: 329–30.

[41] Cf. Saunders Ben. Fairness Between Competing Claims. Res Publica 2010; 16: 42-44.

[42] Broome John. Fairness. Proceedings of the Aristotelian Society; 1990; 91: 87–101 see also: Broome, J. Weighing Goods. Oxford: Blackwell Publishers, 1991:192-200; Broome, J. Kamm on fairness. Philosophy and Phenomenological Research 1998; 58: 955–961; see note 35, Broome 2004:37-40; see note 36, Hooker 2005.

[43] See note 38, Broome 1998: 959.

[44] See note 35, Broome 2004.

[45] Note that this is a simplification. Though claims to medical resources are grounded in medical need, the strength of claims may vary in accordance with other factors such as age. For example, in the current Covid-19 pandemic, an older and a younger person may have the same medical need, in the sense that both have the same probability of survival if put on a ventilator. But it might nevertheless be argued that the younger person has a weaker claim to the resource than the older patient. We are grateful to an anonymous reviewer for pressing us on this point.

Acknowledgements: TG is supported by the Deutsche Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence "Machine Learning—New Perspectives for Science", EXC 2064, project number 390727645). TG also likes to thank Isabel Valera for many helpful discussions on the topic of fairness in machine learning. GK's contribution to this research was funded by the Wellcome Trust project 'Understanding Medical Black Boxes' (Grant Number: 213660/Z/18/Z) during his time at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. It was subsequently funded by the McCoy Family Center for Ethics in Society at Stanford University, via a postdoctoral research fellowship grant from the Stanford Institute for Human-Centered Artificial Intelligence.