AI & SOCIETY
https://doi.org/10.1007/s00146-022-01455-6

ORIGINAL ARTICLE

# Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms

**Benedetta Giovanola**[1,3] · **Simona Tiribelli**[1,2]

Received: 29 August 2021 / Accepted: 13 April 2022


**Abstract**
The increasing implementation of and reliance on machine-learning (ML) algorithms to perform tasks, deliver services and make decisions in health and healthcare have made the need for fairness in ML, and more specifically in healthcare ML algorithms (HMLA), a very important and urgent task. However, while the debate on fairness in the ethics of artificial intelligence (AI) and in HMLA has grown significantly over the last decade, the very concept of fairness as an ethical value has not yet been sufficiently explored. Our paper aims to fill this gap and address the AI ethics principle of fairness from a conceptual standpoint, drawing insights from accounts of fairness elaborated in moral philosophy and using them to conceptualise fairness as an ethical value and to redefine fairness in HMLA accordingly. To achieve our goal, following a first section aimed at clarifying the background, methodology and structure of the paper, in the second section, we provide an overview of the discussion of the AI ethics principle of fairness in HMLA and show that the concept of fairness underlying this debate is framed in purely distributive terms and overlaps with non-discrimination, which is defined in turn as the absence of biases. After showing that this framing is inadequate, in the third section, we pursue an ethical inquiry into the concept of fairness and argue that fairness ought to be conceived of as an ethical value. Following a clarification of the relationship between fairness and non-discrimination, we show that the two do not overlap and that fairness requires much more than just non-discrimination. Moreover, we highlight that fairness not only has a distributive but also a socio-relational dimension. Finally, we pinpoint the constitutive components of fairness. In doing so, we base our arguments on a renewed reflection on the concept of respect, which goes beyond the idea of equal respect to include respect for individual persons. In the fourth section, we analyse the implications of our conceptual redefinition of fairness as an ethical value in the discussion of fairness in HMLA. Here, we claim that fairness requires more than non-discrimination and the absence of biases as well as more than just distribution; it needs to ensure that HMLA respects persons both as persons and as particular individuals. Finally, in the fifth section, we sketch some broader implications and show how our inquiry can contribute to making HMLA and, more generally, AI promote the social good and a fairer society.

**Keywords** Fairness · Healthcare machine-learning algorithms · Bias · Discrimination · Ethics of algorithms · Respect

## 1 Introduction

Fairness is one of the core AI ethics principles and is prominent especially in discussions on machine-learning (ML) algorithms (Mittelstadt et al. 2016; Jobin et al. 2019; Tsamados et al. 2021). In recent years, initiatives focused on fairness in AI have increased greatly, and a growing body of literature has been developed, focusing on the need to address and improve fairness in AI systems (Kleinberg et al. 2017; Edwards and Veale 2017; Overdorf et al. 2018; Binns 2018; Selbst et al. 2019; Wong 2019; Abebe et al. 2020), especially as a response to their controversial effects in a wide array of application domains, including social media

✉ Benedetta Giovanola
benedetta.giovanola@unimc.it

Simona Tiribelli
simona.tiribelli@unimc.it

[1] Department of Political Sciences, Communication, and International Relations, University of Macerata, Macerata 62100, Italy

[2] Present Address: Institute for Technology and Global Health, PathCheck Foundation, 955 Massachusetts Ave, Cambridge, MA 02139, USA

[3] Department of Philosophy, Tufts University, 222 Miner Hall, Medford, MA 02155, USA

Published online: 21 May 2022

communication and information (Bozdag 2013; Shapiro 2000; Hinman 2005, 2008; Laidlaw 2008), advertising and marketing (Hildebrandt 2008; Coll 2013; Tufekci 2015), recruiting and employment (Kim 2017), university admissions (Simonite 2020), housing (Barocas and Selbst, 2016), credit lending (Deville, 2013; Lobosco 2013; Sengh Ah Lee and Floridi 2020), criminal justice (Berk et al. 2018; Abebe et al. 2020) and policing (Ferguson 2017), just to name a few.

The debate on fairness in AI and ML algorithms has expanded significantly in the domain of healthcare as well (Danks and London 2017; Buhmann et al. 2019; Robbins 2019), becoming a progressively urgent topic to tackle, considering the increasing use of algorithmic decision-making in clinical settings and healthcare facilities. In this scenario, the Covid-19 pandemic has made the topic even more urgent: in fact, it has sped up the design and use of novel solutions based on ML for many health-related services, such as health population monitoring (e.g. contact tracing apps), remote assistance and homecare (e.g. telemedicine), and hospitals-and-care access management in response to clinics' overload. This increasing deployment of and reliance on ML algorithms that are trained mainly on patients' personal data to make decisions on health and healthcare have been uncovered as beneficial in terms of health research, operational efficiency, healthcare resources' management and waste reduction, but also as potentially harmful to the promotion of fairness. A recent example of the latter concern is the ML-based software that was applied in thousands of hospitals in the United States of America (US) that runs access to specially resourced care programmes. It has been found to be biased and to reify social disparities, favouring White patients over sicker Black patients in determining patients-in-need priority scores due to an erroneous use of past medical expenditures, which are historically lower among Black patients, as a proxy for determining access to extra medical support (Obermeyer et al. 2019).

However, while the need for fairness in AI and more specifically in healthcare ML algorithms (HMLA) is widely acknowledged as an urgent task (Shin and Park 2019), the very concept of fairness as an ethical value has not been sufficiently explored thus far. Our paper aims to fill this gap and address the AI ethics principle of fairness from a conceptual standpoint, drawing insights from accounts of fairness elaborated in the framework of moral philosophy and using them to conceptualise fairness as an ethical value and redefine fairness in HMLA accordingly.

To achieve our goal, from a methodological standpoint, we conducted a literature review of the concept of fairness underlying both the technical literature on AI and HMLA and the literature in the field of moral philosophy. In this way, we pursued a clarification of the concept of fairness as an ethical value that we consider preliminary to any eventual attempt to integrate fairness—and, more generally, any ethical value or principle—into technological design. In this regard, our inquiry does not overlap with but might be beneficial to the attempts at ethical design developed by safe-by-design (SBD) (Baum 2016) and value-sensitive design (VSD) approaches (van den Hoven, Vermaas and van de Poel 2015; Friedman et al. 2017; Umbrello 2020), especially in the framework of the ethical design of AI for the social good (Umbrello and van de Poel 2021).

The paper is structured as follows. In the second section, we provide an overview of the state of the art of the discussion of the AI ethics principle of fairness in HMLA and show that the concept of fairness underlying this debate is framed in purely distributive terms and overlaps with non-discrimination, which is defined in turn as the absence of biases. At the end of the section, we question whether the concept of fairness so understood is adequate for the discussion of fairness in HMLA or whether the latter calls for a more complex concept of fairness that requires more than just non-discrimination and an exclusively distributive dimension, and that includes features and criteria that extend beyond the consideration of biases.

In the third section, we pursue an ethical inquiry into the concept of fairness and argue that fairness ought to be conceived of as an ethical value. After clarifying the relationship between fairness and non-discrimination, we show that the two do not overlap and that fairness requires much more than just non-discrimination. Moreover, we highlight that fairness not only has a distributive but also a socio-relational dimension. Finally, we pinpoint the constitutive components of fairness. In doing so, we base our arguments on a renewed reflection on the concept of respect, which goes beyond the idea of equal respect to include respect for individual persons.

After unpacking the concept of fairness through our ethical inquiry, in the fourth section, we analyse the implications of our conceptual redefinition of fairness as an ethical value on the discussion of fairness in HMLA. Here, we claim that fairness requires more than non-discrimination and absence of biases as well as more than just distribution; it needs to ensure that HMLA respects persons both as persons and as particular individuals.

In the fifth and final section, we sketch some broader implications of our inquiry and show how it can contribute to making HMLA and, more generally, AI promote the social good and a fairer society.

## 2 Fairness in healthcare machine-learning algorithms

The application of AI and specifically ML algorithms in healthcare has expanded enormously in the last decade (Harerimana et al. 2018; Esteva et al. 2019; Tran et al. 2019).

Similarly, the corpus of literature on the ethics of HMLA is rapidly growing (Morley et al. 2020), making the issue of fairness central (Mittelstadt et al. 2016; Jobin et al. 2019; Shin and Park 2019; Tsamados et al. 2021) and one of the most urgent tasks to tackle today (Grote and Berens 2020; Garattini et al. 2019; Alvin Rajkomar et al. 2018). Fairness in HMLA has been discussed among policymakers, clinical entrepreneurs and computer and data scientists (Grote and Berens 2020; Morley et al. 2020), and it has mostly been understood as the achievement of a state of an absence of biases (Friedler et al. 2016; Char 2018; Obermeyer et al. 2019; Rajkomar et al. 2018).

On the one hand, ML capacity to discover probabilistically correlations, find new patterns and thus produce novel knowledge for health is described as the promise of medicine (Hinton 2018; Norgeot et al. 2019; Chin-Yee and Upshur 2019). It already exhibits great potential in several health application fields, from clinical diagnosis (Álvarez-Machancoses et al. 2019; Fleming 2018), high levels of precision in cancer prediction (Kuo et al. 2001) and diabetes detection (Barakat et al. 2010; Gulshan et al. 2016), to personalised and prevention medicine (Barton et al. 2019), drug discovery (Hay et al. 2013) and epidemiology (Fleming 2018; Álvarez-Machancoses and Fernández-Martínez 2019). Those who advocate for the use of ML in healthcare indeed stress how ML can fix flaws of clinicians, such as their predisposition towards cognitive biases and hence to commit diagnostic errors, and how ML can increase operational efficiency in the healthcare system by reducing resource waste and increasing fairness in access to healthcare services (Topol 2019).

On the other hand, the use of ML algorithms has been discovered to produce highly controversial effects in the domain of healthcare (Danks and London 2017; Buhmann et al. 2019; Robbins 2019), and flaws in healthcare ML systems have recently been denounced. Examples include ML algorithms for heart failures' risk score that have been shown to inappropriately categorise Black patients as being in need of less care (Vyas et al. 2020), as well as ML-based algorithmic models that have been uncovered to be poor at detecting cancers of Black patients (Noor 2020), or to privilege White people over Black patients in the candidates' programme enrolment due to racial biases in the dataset, as it is the case in the US national 'high-risk care management' programme (Obermeyer et al 2019).

These flaws in HMLA, in turn, translate into better or worse opportunities in real access to health quality and health resources as well as extra medical support and economic facilitations for some people over others (Cohen et al. 2014; Morley et al. 2020).

Moreover, these flaws are usually extremely difficult to detect. This is because the ML algorithms used in healthcare are mainly proprietary and, therefore, inscrutable (Burrell 2016). In addition, ML algorithms very often include deep neural networks insofar as they augment the personalisation rate of health prediction. However, because of their complex architectural features, they can easily lead to a lack of transparency and tend to turn into an opaque black box (Pasquale 2015); therefore, auditing and the correction of flaws such as biases and technical inaccuracies is an extremely challenging task to perform, even when ML algorithms' secret trade can be disclosed.

Most of the current scholarship acknowledges the production of unfair outcomes due to biases as one of the main concerns related to ML (Grote and Berens 2020; Morley et al. 2020) and argues that ML systems, rather than simply guarding against these harms passively, should be used proactively to advance ML algorithmic fairness in healthcare (Rajkomar et al. 2018). To this aim, a growing corpus of (mostly technical) literature has been focusing on detecting and fixing biases in HMLA, insofar as biases are deemed to be the main cause of health unfairness (Rajkomar et al. 2018; Char 2018; Obermeyer et al. 2019), and the latter is in turn understood mainly as algorithmic discrimination (Angwin et al. 2016; Hardt et al. 2016).[1]

Algorithmic discrimination is usually understood as the production of discrimination (O'Neil 2016; Noble 2018; Eubanks 2018; Benjamin 2019) in the consideration or treatment of members of protected groups or categories; this discrimination is mostly traced back to the presence of 'automation bias' and 'bias by proxy' in ML algorithmic models. Automation bias is the large-scale spread through ML processes of social and cultural biases that are deeply embedded in the historical training data used to fuel ML algorithms (Hu 2017; Turner Lee, 2018; Noble, 2018; Benjamin, 2019; Richardson et al. 2019; Abebe et al. 2020). Bias by proxy occurs when unanticipated proxies for protected variables (gender, race, etc.) can still be used to reconstruct and infer, by proxy, biases that are highly difficult to detect and eliminate, even though an attempt was made to prevent some biases by excluding them from the historical data used to train the ML model (Fuster et al. 2017; Gillis and Spiess 2019). This operationalisation of fairness as non-discrimination via non-biased ML models is evident when considering the most prominent methods to ensure fairness in algorithmic decision-making and ML, which today consist of 'discrimination prevention analytics and strategies' (Romei and Ruggeri 2014) and 'fairness-and discrimination-aware data mining techniques' (Dwork et al. 2011; Kamishima et al. 2012). These methods are based on the technical engineering

[1] This technical pathway to achieve algorithmic fairness via the development of non-biased ML models also emerges in the wider and more general debate on the ethics of algorithmic decision-making and ML (Barocas 2014; Shah 2018; Diakopoulos and Koliska 2017; Giovanola and Tiribelli 2022).

of anti-discrimination criteria, their integration into the ML classifier and on the control of the distortion of data used to train the algorithms; they also emerge from the main initiatives on fairness in ML in the industry (Ochigame 2019).

In the specific domain of healthcare, four categories of biases have been detected as peculiar to the healthcare ML model (Rajkomar et al. 2018):

(1) Biases in healthcare ML depending on *model design* (such as label biases and cohort biases).

(2) Biases in healthcare ML depending on *training data* (such as minority bias, missing data bias, informativeness bias and training serving skew).

(3) Biases in healthcare ML produced by the ML *interactions with clinicians* (such as automation bias, feedback loops, dismissal bias and allocation discrepancy).

(4) Biases in healthcare ML produced by ML *interactions with patients* (such as privilege bias, informed mistrust and agency bias).

The first category identifies biases that can emerge from the specific design of the healthcare ML algorithmic model (Zafar et al. 2015; Goh et al. 2016; Cotter et al. 2018; Agarwal et al. 2018) and includes, first, *label biases* (Jian and Nachum 2019), which are biases concerning the labelling phase, such as the phase of the selection of reliable annotators, and this kind of bias arises when labels used as proxies are inaccurate; hence, they do not mean the same thing for all patients, leading to the use of an imperfect proxy that is subject to health care disparities rather than an adjudicated truth. Second, there are *cohort biases*; namely, biases depending on the default approach to focus mainly on traditional or easily measured groups without considering other potentially protected groups or levels of granularity (e.g. whether sex is recorded as male, female or other, or more granular categories). Fairness in this sense is achieved by adjusting the ML model rather than the data, despite the fact that frequently it is the training data itself that is biased.

The second category is related to the choice of datasets used to train the model (Tommasi et al. 2015; Angwin et al. 2016; Hardt et al. 2016) and includes the following:

(a) Biases that can emerge from an absence of sufficient representativeness of patients in a protected group for a model to learn the correct statistical pattern (*minority bias*).

(b) Biases depending on a lack of data of patients of protected groups; lack of data that makes an accurate prediction hard to render; for example, a model may under-detect clinical deterioration in patients under contact isolation because they have fewer vital signs (*missing data bias*).

(c) Biases due to the availability of features that are less informative to render an accurate prediction in a protected group; for example, identifying melanoma from an image of a patient with dark skin may be more difficult (*informativeness bias*).

(d) Biases due to the deployment of ML on patients whose data are not similar to the data on which the model was trained (*training–serving skew)*, as in this case, the training data may not be representative (i.e. selection bias), or the deployment data may differ from the training data (e.g. a lack of unified methods for data collection or not recording data with standardised schemas). This is what occurred in the case of Watson for Oncology, a healthcare ML algorithmic system widely used in China for diagnosis prediction via image recognition, which was found to produce poorer results for Chinese patients than their Western counterparts, as the algorithm was primarily trained on Western datasets (Liu et al. 2018).

Fairness in healthcare ML is operationalised in this sense by assessing and re-equilibrating the representativeness of data for protected categories on the basis of which the healthcare ML model is trained and learns to identify patterns, which it uses to produce specific outcomes.

The last two categories identified above instead concern biases that can arise from the interaction between the healthcare ML model with clinicians or patients. In the former (i.e. ML–clinician interaction), it is possible to distinguish the following biases:

(a) *Automation biases* that are due to an overreliance of clinicians on the ML model, also caused by clinicians' unawareness of the inaccuracy of the ML model for a protected group, leading them to act inappropriately on inaccurate predictions.

(b) *Biases of feedback loops* that arise if the clinician accepts the recommendation of a model even when it is incorrect to do so, and the model in turn is so trained to learn mistakes (Mansouri et al. 2020), and this is possible insofar as the model recommended versus administered treatments will always match.

(c) *Dismissal biases* that result from the clinicians' conscious or unconscious desensitisation to alerts that are systematically incorrect for a protected group (e.g. an early warning score for patients with sepsis).

(d) *Biases of allocation discrepancy*, which emerge when some protected groups display disproportionately fewer positive predictions and then resources allocated by the predictions (e.g. extra clinical attention or social services) are withheld from that group.

In the latter case (i.e. ML–patient interaction), it is possible to detect other kinds of biases, which include the following:

(a) *Privilege bias*, i.e. some models may be unavailable in settings where protected groups receive care or require technology/sensors disproportionately available to the nonprotected class, and this also exacerbates existing inequalities between the 'haves' and the 'have-nots' in terms of access to the digital healthcare ecosystem

(Morley et al. 2020); in other words, those that generate enough data on themselves to ensure accurately trained algorithms and those that do not (Topol 2019).

(b) *Informed mistrust bias* that is given by the patients' diffidence based on historical exploitation and unethical practices; protected groups may believe that a model is biased against them, and these patients may avoid seeking care from clinicians or systems that use the model or deliberately omit information, while the protected group may be harmed by this, as it results in them not receiving appropriate care and not interacting with the model, as it enhances the issue of lack of data representativeness and accuracy of that group.

(c) *Agency bias* (deeply connected to privilege bias); protected groups may not have input into the development, use and evaluation of models. Thus, they may not have the resources, education or political influence to detect biases, protest and force correction concerning the consideration or treatment of patients, especially those belonging to protected groups.

Despite some differences, the majority of biases detected in HMLA translate into the re-production of existing discriminations in terms of unequal treatment and consideration of members belonging to protected groups due to HMLA's intrinsic (i.e. model design and training data) and/or relational (i.e. HMLA's interactions with clinicians and patients) features. This is because in the specific corpus of literature still being developed on healthcare and ML, the underlying concept of fairness emerges mainly as an absence of biases leading to discrimination, and discrimination in turn is understood mostly—if not exclusively—as intertwined with distributive issues (Obermeyer et al. 2019; McCradden et al. 2020; Corbett-Davies and Goel, 2018; Benjamin 2019; Friedler et al. 2016; Barocas, 2014; Kleinberg et al. 2017). More specifically, health unfairness is mainly traced back to the presence of biases in HMLA that reflect and, therefore, automate, reify and even exacerbate historical health disparities and discriminations towards protected groups in terms of an unequal distribution of resources (Friedler et al. 2016; Cohen et al. 2014), such as medical care, clinical services and health facilities (including the availability of digital health technology and telemedicine apps).

This underlying emphasis on the distributive dimension of fairness also emerges in the literature on ML, which makes explicit reference to the principles of distributive justice, also called 'distributive justice options' (Rajkomar et al. 2018). Friedler et al. (2016), for example, argue that fairness can be defined as both an 'equality of outcomes' produced by ML and the 'equality of treatment' of different groups of people by the ML model. Rajkomar et al. (2018) further relate to this account to enucleate what they identify as the three distributive justice options to understand fairness in healthcare ML:

(a) *Equal outcomes*, according to which fairness is achievable in ML if the benefits produced from the deployment of ML models in terms of patient outcomes are the same for protected and nonprotected groups.[2]

(b) *Equal performance*, according to which an ML model can be defined as fair if its performance and results are equally accurate for patients in the protected and nonprotected groups for such metrics as accuracy, sensitivity (also known as *equal opportunity*), specificity (or *equalised odds*) and positive predictive value (or *predictive parity*).[3]

(c) *Equal allocation*, also called *demographic parity* (Pleiss et al. 2017), according to which fairness in ML is achieved if the allocation of resources as decided by the model is equal across groups and, especially, proportionally allocated to patients in the protected group. The metric is used to evaluate the rate of positive predictions produced by ML for protected and nonprotected groups.

This understanding of fairness translates into solutions developed to ensure fairness in HMLA that mainly coincide with the creation of neutral or parity models (Corbett-Davies and Goel, 2018); that is, models designed to produce non-discriminatory predictions by constraining biases with respect to predicted outcomes for members of protected groups (Friedler et al. 2016; Corbett-Davies and Goel, 2018). However, this understanding of fairness has been criticised for relying excessively on technical parity and dataset neutrality that is achievable via the elimination of references to protected groups' identities (McCradden et al. 2019) and for not always providing the best techniques to achieve fairness in HMLA. In fact, in the health domain,

[2] According to Rajkomar et al. (2018), a weak form of equal outcomes is ensuring that both the protected and non-protected groups benefit similarly from a model (equal benefit); a stronger form is making sure that both groups benefit and *any* outcome disparity is lessened (equalised outcomes).

[3] Rajkomar et al. (2018) enucleate these metrics with an example that considers African American patients as a protected group: 'A higher false-negative rate in healthcare ML prediction would mean African American patients were missing the opportunity to be identified; in this case, equal sensitivity is desirable. A higher false-positive rate healthcare ML prediction in might be especially deleterious by leading to potentially harmful interventions (such as unnecessary biopsies), motivating equal specificity. When the positive predictive value for alerts in the protected group is lower than in the nonprotected groups, clinicians may learn that the alerts are less informative for them and act on them less (a situation known as class-specific alert fatigue). Ensuring equal positive predictive value is desirable in this case' (p. 5).

the integration of differences between identities is appropriate because there is a reasonable presumption of causation (McCradden et al. 2020), as in the case of biological differences between genders that can affect the efficacy of pharmacological compounds. Omitting variables such as gender and race can diminish accuracy and exacerbate discrimination, instead of mitigating it, while the presence of sensitive attributes can allow HMLA to disclose differently biased correlations, and therefore, help HMLA's designers to develop auditing meta-algorithms that, as they are trained on them, can work for their correction. As in the case of HMLA for health programmes' enrolment (Obermeyer et al. 2019), the removal of sensitive attributes neither prevents the development of biased HMLA nor implies the design of fair HMLA.

Similarly, relying on neutral correlations does not ensure the design of fair HMLA; an example might be HMLA (cleaned by sensitive traits) used by health insurance agencies in the US to determine rates or health coverage for medical expenses. Among the negative factors determining health insurance rates, as well as access to subsidised care programmes, is whether the subject is a smoker. As smoking is recognised as the cause of many diseases, the subject is evaluated by HMLA as being likelier to incur by choice in more medical expenses. On the basis of this correlation, the HMLA may evaluate the subject as not being in financial need, in charge (by choice) of more medical costs, and if ill, with a lower probability of healing from a specific disease. This may cause the subject to experience higher insurance rates or exclude or place her at the bottom of waiting lists for accessing subsidised health programmes, as she is evaluated with no economic priority and, as a smoker, also less entitled to it from both a medical and social standpoint. However, this correlation using a neutral connotator, such as smoking, is not effectively neutral and can reinforce disparities, although cleaned by references to sensitive connotators. Smoking is particularly diffuse, especially in the US, among those living in degrading and harsh socio-economic conditions, where access to health treatments, programmes or insurances is already deeply compromised. As a consequence, this correlation, if fixed, can lead to the design of HMLAs that perpetuate and reinforce, instead of mitigate, existing disparities.

Moreover, the different distributive justice options mentioned above are sometimes incompatible (Friedler et al. 2016; Dieterich et al. 2016); for example, a model may be fair with respect to the outcomes but unfair with respect to the allocation (and vice versa), but it is extremely hard, if not impossible, for any ML model to satisfy all conditions (Chouldechova 2017).

This general difficulty in defining what fairness in HMLA is (Friedler et al. 2016) and how to ensure and promote it emerges in the debate as clear proof of the fact that 'ML fairness is not just a task for ML specialists, but requires ethical reasoning' (Rajkomar et al. 2018) and that 'framing fairness as a purely technical problem is problematic' (McCradden et al. 2020).

In the next section, we attempt to fill this gap and unpack the concept of fairness using ethical reasoning. Through our ethical inquiry on fairness, we aim to highlight important dimensions and components of fairness that have been ignored thus far in the 'purely technical' debate on the topic. This will allow us to adequately evaluate the meaning and role of fairness in HMLA.

## 3 Fairness as an ethical value

As shown in the previous section, fairness in HMLA is mainly understood as an absence of biases and requires the removal of four specific kinds of biases, respectively, on model design, training data, interactions with clinicians and interactions with patients. These biases, in turn, concern HMLA's different treatment and consideration of people belonging to protected and nonprotected groups, leading to discriminating effects. Consequently, the widespread idea is that by detecting and eliminating these biases, it would be possible to mitigate or fix ML algorithmic discrimination (O'Neil, 2016; Noble 2018; Eubanks 2018; Benjamin 2019) and ensure fairness in healthcare ML systems (Rajkomar et al. 2018).

Moreover, the discrimination triggered by HMLA is mostly understood as being intertwined with unfair distribution, while fairness in healthcare ML systems is generally considered to be reached when 'distributive justice options' are taken into account. In particular, an ML model is considered to be fair when its outcomes, performances or effects on patients do not produce discrimination among groups (Rajkomar et al. 2018; Newell and Marabelli 2015; Kleinberg et al. 2017; Corbett-Davies and Goel 2018; Gillis and Spiess 2019). However, as shown above, the different distributive options are often incompatible, and neutral models might not lead to fairness; at the same time, it is highly questionable that fairness requires only a distributive dimension.

In this section, we aim to show that fairness requires more than non-discrimination and does not only have a distributive dimension. Fairness, we claim, is to be understood as an ethical value. To achieve our aim, we first clarify the relationship between fairness and discrimination, drawing on the main reflections offered by moral philosophy. We then zoom in on the ethical significance of fairness and define fairness as an ethical value. In doing so, we base our arguments on a renewed reflection on the concept of respect, which goes beyond the idea of equal respect to include respect for individual persons.

The relationship between fairness and discrimination has been widely acknowledged by philosophical scholarship, mainly in the framework of theories of justice.[4] Many scholars have focused on discrimination as a form of unfair treatment rooted in the misrecognition of the value of equality. The main argument is that every person has equal moral worth, and therefore, deserves equal concern (Dworkin 2000). This implies treating people as equals and refraining from wrongful discrimination,[5] as far as distributive justice is concerned.

Other scholars have focused on the social meaning of discrimination. From this perspective, discrimination is seen as a way of demeaning or degrading someone and implies treating them cruelly or humiliating them to undermine their capacity to develop and maintain an integral sense of self (Sangiovanni 2017). In a similar vein, other scholars, drawing on Rawls's justice as fairness (1971), focus on the socio-relational aspects of discrimination and highlight its negative effects on the achievement of a society of equals. In this view, discrimination is considered a major moral and social wrong, as it hinders attitudes and practices of mutual recognition among persons (Scheffler 2003; Anderson 1999).

Despite the acknowledgement of the moral value of equality for discussions on discrimination and fairness, relatively little work has been done so far on the ethical significance of discrimination and fairness themselves as well as on the ethical import of their relationship.

As for *discrimination*, only a few recent studies have focused on the moral wrong of discrimination and investigated the conditions under which discrimination is wrongful (Eidelson 2015; Lipper-Rasmussen 2013; Moreau 2010). The general idea underpinning these works is that wrongful discrimination is connected to moral disrespect—that is, disrespect for the discriminatees as persons (Eidelson 2015, p. 6). More precisely, an action is discriminatory if either the reasons underlying the action or the consequences brought about by the action do not respect the status of an agent as equal. In other words, it is 'the absence of appropriate responsiveness to someone's standing as a person' (Eidelson 2015, p. 7) that underpins moral disrespect and wrongful discrimination. The moral respect at stake here can best be captured by referring to the notion of recognition respect elaborated by Darwall (1977): respect grounded in the recognition of the (equal) humanity of every person.[6]

Highlighting the ethical significance of discrimination helps shed light on the ethical import of the relationship between discrimination and fairness: discrimination emerges as a moral wrong that shows disrespect for people in as far as it denies their standing as persons; that is, their moral equality. Denying someone's moral equality prevents mutual recognition as equals and fosters unequal treatment. It follows that the moral wrong of discrimination can prevent fairness in at least two ways: first, by having an impact at a socio-relational level, in that it creates a society of unequals, where people do not respect one another as (equal) persons, as they do not recognise each other as (equal) persons; second, by having an impact at a distributive level, by legitimising an unfair distribution[7] that can become structural.[8]

As for *fairness*, ethical inquiry has mainly been aimed at investigating the basis of moral equality that fairness ought to ensure (Waldron 2017; Carter 2011; Sangiovanni 2017), rather than at discussing the ethical significance of fairness as such. A widespread idea is that moral equality can be guaranteed through fair distribution; the latter, in turn, would require compliance with the principle of 'fair equality of opportunity' and a 'difference principle' (Rawls 1971). Both principles regulate the distribution of benefits and burdens of social cooperation, but while the former prescribes arranging socio-economic inequalities in such ways that are not influenced by morally arbitrary factors and rather enable everyone's personal agency and self-realisation (Rawls 1971, p. 73), the latter requires that—once the former principle is guaranteed—the overall scheme of cooperation and distribution does not discriminate against the (expectations of the) worst off.

The emphasis on the distributive dimension of fairness has led some scholars—labelled as luck egalitarians

[4] Foundational questions about discrimination are familiar to legal scholars, too, and in recent years, in particular, there has been a renewed interest in philosophical questions about anti-discrimination law (Khaitan 2015; Hellman and Moreau 2013) aimed mainly at defining under what conditions discrimination ought to be prohibited. The focus of these inquiries, however, is on discrimination rather than on the relationship between discrimination and fairness.

[5] However, drawing on Dworkin (2000), Waldron (2017, p. 14) acknowledges that not every discrimination is wrongful; in fact, there might also be forms of unequal treatment or 'surface-level' discrimination that do not imply any moral wrongdoing, but rather are justifiable by an appeal to the whole range of human interests, as in the case, discussed by Waldron, of firefighters being selected for their physical fitness.

[6] Darwall (1977) introduces the well-known distinction between *recognition respect* and *appraisal respect*, whereby the latter depends on the appraisal of a person's character. Darwall's account of recognition respect has been further elaborated on by Carter (2011), who develops the notion of 'opacity-respect'; that is, recognition respect expressed through the idea that we have to treat every person as 'opaque', respecting them on the footing of moral equality, without engaging in an assessment of their personal merits or demerits (Carter 2011).

[7] The question remains open regarding what ought to be distributed, such as with resources, opportunities or outcomes. This is the well-known issue of the 'metrics' or 'currency' of justice: for an overview, see Brighouse and Robeyns (2010).

[8] Just to mention one of the most problematic cases, consider racial injustice; it is historically rooted, socially shaped and institutionally entrenched in distributive policies: see Shelby (2016), Kelly (2017).

(Anderson 1999)—to claim that fairness requires treating people as equals by considering individual choices and their effects on distribution. In this view, distribution, to be fair, ought to be 'ambition-sensitive' (or choice- or responsibility-sensitive) and 'endowment-insensitive'. The former condition requires that distribution depend on choices made by individuals in ways that reflect their option luck, whereas the latter condition requires that distribution does not depend on differential brute luck. Brute luck is fortune, over which individuals have no control, while option luck is the upshot of risks that were, in some sense, deliberately taken and for which individuals have responsibility.[9] However, luck egalitarian accounts have been criticised for being too harsh in as far as they would eventually lead to abandoning the so-called 'negligent victims', or those who would end up in a bad situation due to the risks they had deliberately taken. This abandonment, it has been claimed, would be deeply disrespectful of persons (Anderson 1999; Scheffler 2003; Wolff 1998, 2010), as it would ultimately clash with the acknowledgement of every person's moral equality. Moreover, it would not only have a discriminating impact at the distributive level towards the 'negligent victims', but it would also bring about socio-relational inequalities, as the 'negligent victims' would be looked at as less valuable citizens, not deserving equal respect, due to the choices they made.

The latter considerations confirm that fairness goes beyond distribution and entails a socio-relational dimension (Giovanola 2018). The underlying idea here is that moral equality can only be guaranteed if people stand in a relation of equality and recognise one another as equals (Anderson 1999; Scheffler 2003; Wolff 1998, 2010). In this view, as already noted, fairness is aimed at the creation of a society of equals where no wrongful discrimination occurs. However, if we dig deeper into the socio-relational dimension of fairness, we find that it requires more than this kind of non-discrimination: it also requires guaranteeing everyone an equal 'right to justification' (Forst 2014). The right to justification expresses the ethical demand that no 'relations should exist that cannot be adequately justified towards those involved' (Forst 2014, p. 6); it points to the need for intersubjective relations and structures that do not discriminate against anyone, protect every person's status of an equal and express (equal) respect for the equal moral worth of every person.

Like the principle of fair equality of opportunity and the difference principle, the right to justification is grounded in the need to *respect persons as persons;* however, instead of focusing on distribution, it focuses on intersubjective relations and structures and requires that they protect every person's status and capability to make up their own minds on issues of concern. This demand rests on a principle of general and reciprocal justification; that is, on the claim that every person ought to be respected as a subject who offers and demands justification. Therefore, the question of justification is also a question of power or the question of who decides what (Forst 2014, p. 24).

To summarise our arguments thus far, fairness requires compliance with 'fair equality of opportunity' and a 'difference principle' as well as an equal 'right to justification'. While the first two components highlight the distributive dimension of fairness, the third uncovers its socio-relational dimension. All components are grounded in the recognition of the equal moral worth of each person, which in turn calls for equal respect as the appropriate ethical response to each individual's standing as a person. It follows that fairness is an ethical value of intrinsic moral importance that requires a commitment to ensure *equal respect for persons as persons*.

At this point, however, it should be noted that people's moral worth is not only attached to their status as persons 'but also to their status as particular individuals' (Noggle 1999, p. 457) who exercise their agency in different concrete ways. Acknowledging this implies going beyond an exclusive focus on equal respect to investigate *respect for persons as particular individuals* or particular agents, taking into account the different ways in which different individuals exercise their agency.

A helpful clue for digging into this issue is provided by Noggle, who argues that 'a person is much more than a mere instance of rational agency. She is a being with a particular life, a particular psychology, and a particular set of attachments, goals and commitments. To be a person is not merely to be an instance of rational agency; it is also to be a *particular* individual. It seems that if we are truly serious about respecting persons, we ought to respect them not only as instances of rational agency, but also as the particular individuals that they are' (Noggle 1999, p. 454). A person's particular identity—that is, their status as the particular person that they are—depends on many factors, including their ends, values, attachments, commitments and relations, that make them a concrete 'me', as opposed to a 'disencumbered' and abstract self (Sandel 1984).

Therefore, respecting persons requires respecting their status as particular individuals, too, going beyond treating them as opaque only and recognising the importance of the 'ground projects' that give meaning and purpose to their lives (Williams 1981). It also involves focusing on the

[9] The luck egalitarian account of fairness has been widely applied to the domain of health and healthcare, arguing that people's health and the health care they receive are just when the effects of bad luck only are neutralised (Segall, 2010). For an alternative proposal, drawing insights from Rawls's principle of fair equality of opportunity, see Daniels (1985).

different ways in which they exercise their agency as well as on the ways in which social relations and interpersonal relationships affect their agency and their capability to set ends.[10]

Finally, acknowledging the importance of respect for particular individuals allows us to take a step forward in our conceptual analysis of fairness. While fair equality of opportunity, the difference principle and the right to justification are grounded in the acknowledgement of the equal moral worth of every person and require a commitment to ensure *equal respect for persons as persons*, their effective implementation depends on many factors that affect people's conditions and prospects in life. They also have an impact on their capability to make up their own minds as well as on the opportunities that they have, in ways that vary from individual to individual. These differences trigger demands of respect; that is, *respect for persons as particular individuals*.

The concept of fairness emerging from our inquiry sheds light on fairness as an ethical value. Understanding fairness as an ethical value amounts to redefining fairness, going beyond an exclusive focus on discrimination and accounting for both a distributive and socio-relational dimension based on the acknowledgement of the importance of respect—both for persons as persons and for particular individuals.

Having unpacked and redefined the concept of fairness through our ethical inquiry, in the next section, we analyse its implications for the discussion on fairness in HMLA.

## 4 Fairness in HMLA revised

In Sect. 2, we claimed that fairness in HMLA emerges mainly as non-discrimination and is conceptualised almost exclusively in distributive terms: a fair HMLA is a system capable of ensuring equal treatment (in performance, outcomes and HMLA benefits' distribution) of members belonging to protected and not protected groups via the elimination of four families of biases. We also questioned whether the definition of a fair HMLA as bias-free HMLA emerging from the literature is enough to promote fairness via AI, specifically HMLA in healthcare, or whether a more complex account of fairness is needed. In Sect. 3, we showed that an ethical inquiry into the concept of fairness can help us to adequately elaborate the ethics principle of fairness and clarify that fairness neither overlaps with non-discrimination nor only has a distributive component but rather entails the consideration of both a distributive and a socio-relational dimension. Specifically, we argued that fairness is an ethical value entailing three components, all grounded in the need to respect persons both as persons and as particular individuals: *fair equality of opportunity*, *difference principle* and *equal right to justification*.

In this section, we analyse the implications of our redefinition of fairness as an ethical value on the discussion of fairness in HMLA and highlight areas where (and why) further work is needed, especially at the technical and policy level, to ensure fairness via AI (and specifically ML) in the health sector. More specifically, we show that the three components of fairness are crucial to further develop the ethics principle of fairness in HMLA and to identify what is needed to promote a fairer digital health ecosystem.

As shown in the previous section, *fair equality of opportunity* and the *difference principle* specify the conditions that ought to be met to promote a fair distribution of resources and opportunities via HMLA.

The concept of *fair equality of opportunity* underlies the discussion on fairness in HMLA—as well as the broader debate on AI and ML (Hardt et al. 2016)—and emerges mainly as the inspiring distributive justice tenet to operationalise fairness via HMLA (Rajkomar et al. 2018; Friedler et al. 2016). However, as we have shown, it has been understood and used only in a narrow sense, as being aimed at preventing discrimination through distribution; the underlying idea is that non-discrimination is ensured by HMLA if HMLA guarantee an equal allocation of outcomes, performances or resources (i.e. demographic parity) for protected and nonprotected groups. Moreover, this condition is satisfied via solutions mainly coinciding with parity models (Hardt et al. 2019; Corbett-Davies and Goel, 2018; Rajkomar et al. 2018; Friedler et al. 2016) designed to produce non-discriminatory predictions by eliminating references to protected groups' identities, such as sensitive attributes like race or gender (McCradden et al. 2020). However, as argued in the previous sections, the focus on non-discrimination, pursued through the removal of sensitive attributes and the pursuit of neutral or parity models, does not lead to fairness in HMLA and can sometimes even perpetuate and reinforce wrongful discrimination and existing disparities. However, if the removal of specific attributes and the use of neutral models per se do not promote fair equality of opportunity, what does fair equality of opportunity require when applied to promote fairness in HMLA?

Fair equality of opportunity entails that the distribution of shares—and, more specifically, access to opportunities—is not improperly influenced by socio-economic contingencies, namely by a person's place in the social system (Rawls 1971, p. 63). It does not only require a formal equality of opportunity, ensured, for example, through the legal system; it entails the substantial promotion of opportunities as real chances for every person to express their agency and,

[10] For a more detailed reflection on respect for particular individuals, social relations and interpersonal relationships, see Giovanola and Sala 2021; for an inquiry into the ways in which technology impacts on social relations and interpersonal relationships, as well as on individual's agency and sense for justice, see Giovanola 2021.

therefore, the development of adequate conditions for people to afford them. For this reason, fair equality of opportunity needs to be considered along with the *difference principle,* which emphasises attention to the expectations and conditions of the least advantaged, not only to access but also to substantially enjoy those chances.

Therefore, ensuring fair equality of opportunity and the difference principle in HMLA requires the design of compensatory tools that rather than only adjusting the model (e.g. fixing pernicious biases, that can be performed by auditing meta-algorithms) are thought and used to mitigate social disparities and individuals' capacity to enjoy opportunities, with specific attention paid to the least advantaged. It is now clearer why neither the development of neutral and parity models for HMLA nor the ML's labelling or categorisation of people in groups on the basis of standard or macro-generalised attributes can foster fairness; rather, they can obscure some crucial differences, whose non-consideration is critical from both a medical and social standpoint and can deeply undermine the respect for persons, both as persons and as particular individuals, who as such are situated in deeply different socio-economic contexts. Fair equality of opportunity and the difference principle require that HMLA are informed with sensitive traits (gender, age and race) to the extent that they can play as decisional nodes to evaluate who in a certain health domain may require compensatory tools (i.e. the implementation of extra health measures, support or facilities) if explicitly requested by the subject. The 'explicitly' clause is crucial, as not every person, although a member of a protected group *per antonomasia,* would really need or be willing to be defined as such. For example, an elderly person may require further assistance to access a programme whose application procedure is only online, or she may not; in both cases, the person, ignored in her need or considered vulnerable when she is not, if unheard, can feel differently treated and not properly recognised as the particular individual she is. For this reason, as we will argue below, HMLA should also inform the patient, subject to their decision about the profile or label they have been assigned (e.g. if they are evaluated as belonging to a certain group and why), and consequently intervene, displaying compensatory options when explicitly asked or properly recognised by HMLA. In this way, HMLA would put the person in the position to participate in the specific modelling and alignment of the profile assigned to her with the particular individual (with specific needs, features and ground project) that she is.

To ensure fair equality of opportunity and the difference principle in HMLA, and therefore, to promote real chances for every person to enjoy the digital health ecosystem's resources and facilities equally, researchers and engineers are called to work on methods focusing on both extending these sensitive attributes (also to health issues, e.g. mental health issues) and making them reliable connotators for compensatory tools, which can intervene through the 'explicitly' clause. They can also intervene via, for example, not-profilable semi-structured interviews, while implementing at the same time meta-algorithms capable of detecting and fixing pernicious biases reflecting historical inequalities arising from them.

Moreover, as argued in Sect. 3, fairness as an ethical value entails a socio-relational dimension; this means that an *equal right to justification* is the third condition that ought to be ensured and implemented by a fair HMLA. The equal right to justification expresses the ethical demand that every person be respected as a subject who can offer and demand justification. As introduced via the formulation of the 'explicitly' clause, to promote fairness via HMLA, individuals should be recognised as the particular individuals that they are, with specific needs, vulnerabilities and ground projects. This is fundamental not only for the promotion of a society of equals regarding the distribution of opportunities, but also for the creation of a society of individuals who feel recognised in their voices and actions and are able to effectively understand how their chances and choices are influenced by HMLA. In particular, the right to justification requires that every person has the right to demand justification for the HMLA treatment (from decision to prediction) she is subjected to, not just in output but also in input and process. Therefore, ensuring an equal right to justification in HMLA entails that designers have a duty to take this demand into account in ways that are accessible to the subjects involved. Respecting the equal right to justification does not imply full transparency, as the latter is often unnecessary for a sufficient or adequate explanation to users; rather, it requires informing the patients about information as input and correlations as reasons leading to the development of a certain profile of them on the basis of which persons are subjected to certain decisions, such as having access or not to certain health facilities or a more or less high health insurance rate. This is of particular importance for at least two reasons. First, it would allow patients to ask for reasons for and eventually contest the outcomes/decisions they are subjected to and to ask for changes. This is crucial, as the current opacity characterising HMLA's functioning and, therefore patients' treatment constitutes an asymmetry of power, both in knowledge and in action between those that can be wrongly considered and those who are behind the HMLAs' design, use and deployment. Therefore, ensuring an equal right to justification would effectively mitigate asymmetries of power related to the question of who decides what. Second, it would allow patients to act against the epistemic injustice[11] that HMLA can produce. Indeed, by undermining

[11] The issue of epistemic injustice was first extensively discussed by Fricker (2007), who distinguishes two forms: testimonial injustice and hermeneutical injustice. The former occurs when a speaker

the subject in knowledge, those who are unfairly treated are undermined in their position to participate in the modelling of HMLA as novel health determinants. As HMLA are increasingly applied to manage crucial tasks and services in healthcare, the silencing and exclusion of segments of populations from contesting and asking for revisions of HMLA means hindering fairness in as far as it concerns the promotion of a society of equals, where everyone ought to have an equal opportunity to participate in the processes of defining the (technological) structures that shape society.

To ensure an equal right to justification, HMLA's designers and policymakers are called into action. The former are mobilised in the development of methods that can make HMLA's functioning intelligible to the patients while not fully disclosing it or infringing the company's intellectual property rights; the latter in a) discerning when they are called to adopt HMLA in crucial social sectors, such as healthcare, as the best deployment option to comply with fairness as an ethical value; b) prohibiting an HMLA when a person's request for justification cannot be fulfilled and c) providing the social support (structures and assistance) when a contestation is claimed by patients against HMLAs' outcomes.

In short, redefining the AI ethics principle of fairness in HMLA requires identifying and implementing ways that allow the respecting of persons both as persons and as particular individuals by ensuring that HMLA are designed and applied in compliance with fair equality of opportunity, the difference principle and the equal right to justification.

The question, however, remains open as to whether and how the implementation and deployment of fair HMLA might vary depending on different contexts, both in geographical and political-economic terms.[12] In fact, HMLA comes into play in situations that are already shaped both politically and economically; these situations, in turn, are different in different areas of the world or in different historical periods. For example, distribution-related questions might be more pressing in countries that have more progressive welfare states, and there they would be tackled at the policy rather than business or provider level; this would imply that people expect institutions rather than the market to distribute access to healthcare fairly and that institutions—rather than the market—are held responsible for such a distribution and for the tools and methods of distribution to be fair. At the same time, people's real opportunities to access the distribution of healthcare fairly are strongly influenced not only by the political or economic setting but also by a wide range of factors that are socially or historically determined. In this regard, uncovering discrimination and biases in HMLA, making people aware of these discrimination and biases, and designing and implementing fair HMLA in ways that include those suggested in our paper may contribute to shedding light not only on the conditions for a fair distribution of healthcare through HMLA but also on the threats and opportunities of any distributive tool or methods.

## 5 Concluding remarks

In our paper, we tackled one of the most urgent risks of AI systems in healthcare: the risk of unfairness. In pursuing our analysis, we focused on HMLA and discussed the concept of fairness that is emerging in the debate. We highlighted that fairness in HMLA is mostly framed in distributive terms and overlaps with non-discrimination, which is defined in turn as the absence of biases. We questioned such a concept of fairness and maintained that fairness requires more than the removal of biases and the development of non-discrimination techniques.

Drawing insights from moral philosophy, we proposed a more complex account of fairness as an ethical value based on a renewed reflection on the concept of respect, which goes beyond the idea of equal respect for persons to include respect for particular individuals. In particular, we argued that fairness as an ethical value has both a distributive and a socio-relational dimension and comprises three components: fair equality of opportunity, difference principle and equal right to justification. Finally, we analysed the implications of our conceptual redefinition of fairness as an ethical value in the discussion of fairness in HMLA and highlighted specific areas where further work needs to be done to operationalise fairness in HMLA. We claimed that an ethically informed principle of fairness requires ensuring that HMLA are designed and applied in compliance with fair equality of opportunity, the difference principle and the equal right to justification in ways that respect persons both as persons and as particular individuals and that acknowledge both the distributive and the socio-relational dimension of fairness. By doing so, we invite future research to not only focus on anti-discrimination and bias removal techniques but also to develop novel technical and policy-oriented tools and methods that can promote the fundamental components and dimensions of fairness, and therefore, effectively implement the AI ethics principle of fairness in HMLA.

The revision of the AI ethics principle of fairness in HMLA proposed in our paper indeed shows that HMLA can contribute not only to a fairer healthcare ecosystem but

Footnote 11 (continued)

is given less credibility than deserved because of an identity prejudice held by the hearer; to suffer a credibility deficit in turn impedes one's capacity as an epistemic agent, making it both an ethical and an epistemic wrong. The latter occurs when there exists a lack of collective interpretative resources required for a group to understand (and express) significant aspects of their social experience.

[12] We thank an anonymous reviewer for pushing us to make these issues explicit.

also to the promotion of a fairer society. In as far as they contribute both to a fairer distribution of opportunities for all, and especially for the worst off, to the creation of a society of equals and to respecting every person, HMLA contributes to the social good of society. The social good entails that every person is recognised as equal and given fair opportunities and effective power in knowledge and action. Complying with the revised AI ethics principle of fairness, HMLA can, therefore, become a model for the promotion of a fairer and more inclusive society, where AI can truly become a force for good.

## References


Abebe R, Barocas S, Kleinberg J, Levy K, Raghavan M, Robinson DG (2020) Roles for computing in social change. https://doi.org/10.1145/3351095.3372871. ArXiv:1912.04883.

Agarwal A, Beygelzimer A, Dudik M, Langford J., Wallach H (2018) A reductions approach to fair classification. In: Proceedings of the 35th International Conference on Machine Learning. In Proceedings of Machine Learning Research, 80: 60–69. Available at https://proceedings.mlr.press/v80/agarwal18a.html

Álvarez-Machancoses Ó, Fernández-Martínez JL (2019) Using artificial intelligence methods to speed up drug discovery. Expert Opin Drug Discov 14(8):769–777. https://doi.org/10.1080/17460441.2019.1621284

Anderson E (1999) What is the point of equality? Ethics 109(2):289–337. https://doi.org/10.1086/233897

Angwin J, Larson J, Mattu S, Lauren K (2016) Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Retrieved March 10, 2021

Barakat N, Bradley AP, Barakat MNH (2010) Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed 14(4):1114–1120. https://doi.org/10.1109/TITB.2009.2039485

Barocas S (2014) Data mining and the discourse on discrimination. In: Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining (KDD). https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf. Retrieved March 10 2021

Barocas S, Selbst AD (2016) Big data's disparate impact. SSRN Electron J. https://doi.org/10.2139/ssrn.2477899

Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, Calvert J, Das R (2019) Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 109:79–84. https://doi.org/10.1016/j.compbiomed.2019.04.027

Baum SD (2016) On the promotion of safe and socially beneficial artificial intelligence. AI Soc. https://doi.org/10.1007/s00146-016-0677-0

Benjamin R (2019) Race after technology: abolitionist tools for the new jim code. Polity, Medford

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res. https://doi.org/10.1177/0049124118782533

Binns R (2018) Fairness in machine learning: lessons from political philosophy. http://arxiv.org/abs/1712.03586. Retrieved 11 March, 2021

Bozdag E (2013) Bias in algorithmic filtering and personalization. Ethics Inf Technol 15:209–227. https://doi.org/10.1007/s10676-013-9321-6

Buhmann A, Paßmann J, Fieseler C (2019) Managing algorithmic accountability: balancing reputational concerns, engagement strategies, and the potential of rational discourse. J Bus Ethics. https://doi.org/10.1007/s10551-019-04226-4

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512

Brighouse H, Robeyns I (2010) Measuring justice. Primary Goods and capabilities. Cambridge University Press, Cambridge

Carter I (2011) Respect and the basis of equality. Ethics 121(3):538–571. https://doi.org/10.1086/658897

Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 378(11):981–983

Chin-Yee B, Upshur R (2019) Three problems with big data and artificial intelligence in medicine. Perspect Biol Med 62(2):237–256. https://doi.org/10.1353/pbm.2019.0012

Cohen IG, Amarasingham R, Shah A, Xie B, Lo B (2014) The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff 33(7):1139–1147. https://doi.org/10.1377/hlthaff.2014.0048

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163. https://doi.org/10.1089/big.2016.0047

Coll S (2013) Consumption as biopower: governing bodies with loyalty cards. J Consu Cult 13(3):201–220. https://doi.org/10.1177/1469540513480159

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: a critical review of fair machine learning. http://arxiv.org/abs/1808.00023. Retrieved March 11, 2021

Cotter A, Jiang H, Sridharan K (2018) Two-player games for efficient non-convex constrained optimization. arXiv preprint arXiv:1804.06500.

Daniels N (1985) Just health care. Cambridge University Press, Cambridge

Danks D, London AJ (2017) Algorithmic bias in autonomous systems. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, pp 4691–4697. https://doi.org/10.24963/ijcai.2017/654.

Darwall S (1977) Two kinds of respect. Ethics 88:36–49. https://doi.org/10.1086/292054

Deville J (2013) Leaky Data: How Wonga Makes Lending decisions. Charisma: Consumer Market Studies. http://www.charisma-network.net/finance/leaky-data-how-wonga-makes-lending-decisions. Retrieved March 11, 2021

Diakopoulos N, Koliska M (2017) Algorithmic transparency in the news media. Digit J 5(7):809–828. https://doi.org/10.1080/21670811.2016.1208053

Dieterich B, Mendoza C., Brennan T (2016) COMPAS risk scales: demonstrating accuracy equity and predictive parity performance of the COMPAS risk scales in broward county. https://www.semanticscholar.org/paper/COMPAS-Risk-Scales-%3A-Demonstrating-Accuracy-Equity/cb6a2c110f9fe675799c6aefe1082bb6390fdf49. Retrieved March 11, 2021

Dwork C, Hard M, Pitassi T, Reingold O, Zemel R (2011) Fairness through awareness. http://arxi-v.org/abs/1104.3913. Retrieved March 11, 2021

Dworkin R (2000) Sovereign virtue: the theory and practice of equality. Harvard University Press, Cambridge

Edwards L, Veale M (2017) Slave to the algorithm? Why a right to explanationn is probably not the remedy you are looking for. SSRN Electron J. https://doi.org/10.2139/ssrn.2972855.

Eidelson B (2015) Discrimination and disrespect. Oxford University Press, Oxford

Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J (2019) A guide to deep learning in healthcare. Nat Med 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z

Eubanks V (2018) Automating inequality. How high-tech tools profile, police, and punish the poor. St Martin's Publishing, New York

Ferguson AG (2017) The rise of big dtata policing. Surveillance, race, and the future of law enforcement. New York University Press, New York

Fleming N (2018) How artificial intelligence is changing drug discovery. Nature 557(7707):S55–S57. https://doi.org/10.1038/d41586-018-05267-x

Forst R (2014) Two pictures of justice. In: Justice, Democracy and the Right to Justification. Rainer Forst in Dialogue, Bloomsbury, London, pp 3–26.

Fricker M (2007) Epistemic injustice: power and the ethics of knowing. Oxford University Press, New York

Friedler S, Scheidegger C, Venkatasubramanian S (2016) On the (im) possibility of fairness. https://www.researchgate.net/publication/308610093_On_the_impossibility_of_fairness/citation/download. Retrieved March 11, 2021

Friedman B, Hendry DG, Borning A (2017) A survey of value sensitive design methods. Foundations and Trends®. Human Comput Interact 11(2):63–125. https://doi.org/10.1561/110000001

Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2017) Predictably unequal? The effects of machine learning on credit markets. SSRN Electron J. https://doi.org/10.2139/ssrn.3072038.

Garattini C, Raffle J, Aisyah DN, Sartain F, Kozlakidis Z (2019) Big data analytics, infectious diseases and associated ethical impacts. Philos Technol 32(1):69–85. https://doi.org/10.1007/s13347-017-0278-y

Gillis TB, Spiess J (2019) Big data and discrimination. Univ Chicago Law Rev. https://lawreview.uchicago.edu/sites/lawreview.uchicago.edu/files/09%20Gillis%20%26%20Spiess_SYMP_Post-SA%20%28BE%29.pdf. Retrieved March 11, 2021

Giovanola B (2018) Giustizia sociale. Eguaglianza e rispetto nelle società diseguali. Il Mulino, Bologna.

Giovanola B (2021) Justice, emotions, socially disruptive technologies. Crit Rev Int Soc Polit Philos. https://doi.org/10.1080/13698230.2021.1893255

Giovanola B, Sala R (2021) The reasons of the unreasonable: is political liberalism still an option? Philos Soc Crit. https://doi.org/10.1177/01914537211040568

Giovanola B, Tiribelli S (2022) Weapons of Moral construction? On the value of fairness in algorithmic decision-making. Ethics Inform Technol. https://doi.org/10.1007/s10676-022-09622-5

Goh G, Cotter A, Gupta M, Friedlander MP (2016) Satisfying real-world goals with dataset con- straints. In: Advances in Neural Information Processing Systems, pp 2415–2423. Available at: https://papers.nips.cc/paper/2016/file/dc4c44f624d600aa568390f1f1104aa0-Paper.pdf

Grote T, Berens P (2020) On the ethics of algorithmic decision-making in healthcare. J Med Ethics 46(3):205–211. https://doi.org/10.1136/medethics-2019-105586

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22):2402–2410. https://doi.org/10.1001/jama.2016.17216

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. https://arxiv.org/abs/1610.02413. Retrieved March 12, 2021

Harerimana G, Jang B, Kim JW, Park HK (2018) Health big data analytics: a technology survey. IEEE Access 6:65661–65678. https://doi.org/10.1109/ACCESS.2018.2878254

Hellman D, Moreau S (2013) Philosophical foundations of discrimination law. Oxford University Press, Oxford

Hildebrandt M (2008) Defining profiling: a new type of knowledge?. In: Hildebrandt M, Gutwirth S (eds) Profiling the European Citizen. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6914-7_2

Hinman LM (2005) Esse est indicato in Google: Ethical and Political Issues in Search Engines. International Review of Information Ethics 3. Retrieved March 11, 2021, from https://informationethics.ca/index.php/irie/article/view/345.

Hinman LM (2008) Searching ethics: the role of search engines in the construction and distribution of knowledge. In: Spink A, Zimmer M (eds) Web search. Information science and knowledge management, Springer. https://doi.org/10.1007/978-3-540-75829-7_5.

Hay SI, George DB, Moyes CL, Brownstein JS (2013) Big data opportunities for global infectious disease surveillance. PLoS Med 10(4):e1001413. https://doi.org/10.1371/journal.pmed.1001413

Hinton G (2018) Deep learning-a technology with the potential to transform health care. JAMA 320(11):1101–1102. https://doi.org/10.1001/jama.2018.11100

Hu M (2017) Algorithmic jim crow. Fordham Law Rev. https://ir.lawnet.fordham.edu/flr/vol86/iss2/13/. Retrieved March 10, 2021

Jobin A, Ienca M, Vayena E (2019) Artificial intelligence: the global landscape of ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Considerations on fairness-aware data mining. In: IEEE 12th International Conference on Data Mining Workshops, Brussels, Belgium, pp 378–385. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6406465. Retrieved March 10, 2021

Kelly E (2017) The historical injustice problem for political liberalism. Ethics 128:75–94

Kim PT (2017) Data-driven discrimination at work. 58 Wm. & Mary L. Rev 857(3). https://scholarship.law.wm.edu/wmlr/vol58/iss3/4. Retrieved March 11, 2021

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions. Q J Econ. https://doi.org/10.1093/qje/qjx032

Khaitan T (2015) A theory of discrimination law. Oxford University Press, Oxford

Kuo WJ, Chang RF, Chen DR, Lee CC (2001) Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast Cancer Res Treat 66(1):51–57. https://doi.org/10.1023/A:1010676701382

Laidlaw EB (2008) Private power, public interest: an examination of search engine accountability. Int J Law Inform Technol 17(1):113–145. https://doi.org/10.1093/ijlit/ean018

Lippert-Rasmussen K (2013) Born free and equal? A philosophical inquiry into the nature of discrimination. Oxford University Press, Oxford

Lobosco K (2013) Facebook friends could change your credit score. CNN Business. https://money.cnn.com/2013/08/26/technology/social/facebook-credit-score/index.html. . Retrieved March 11, 2021

Mansoury M, Abdollahpouri H, Pechenizkiy M, Mobasher B, Burke R (2020) Feedback loop and bias amplification in recommender systems. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA: 2145–2148. https://doi.org/10.1145/3340531.3412152.

McCradden MD, Joshi S, Mazwi M, Anderson JA (2020) Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digital Health 2(5):e221–e223. https://doi.org/10.1016/S2589-7500(20)30065-0

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679

Newell S, Marabelli M (2015) Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of 'datificaion.' J Strateg Inf Syst 24(1):3–14. https://doi.org/10.1016/j.jsis.2015.02.001

Moreau S (2010) What is discrimination? Philos Public Aff 38(2):143–179. https://doi.org/10.1111/j.1088-4963.2010.01181.x

Morley J, Machado C, Burr C, Cowls J, Joshi I, Taddeo M, Floridi L (2020) The ethics of AI in health care: a mapping review. Soc Sci Med 260:113172. https://doi.org/10.1016/j.socscimed.2020.113172

Noble SU (2018) Algorithms of oppression: how search engines reinforce racism. New York University Press, New York

Noggle R (1999) Kantian respect and particular persons. Can J Philos 29:449–477. https://doi.org/10.1080/00455091.1999.10717521

Noor P (2020) Can we trust AI not to further embed racial bias and prejudice? BMJ (Clin Res Ed) 368:m363. https://doi.org/10.1136/bmj.m363

Norgeot B, Glicksberg BS, Butte AJ (2019) A call for deep-learning healthcare. Nat Med 25(1):14–15. https://doi.org/10.1038/s41591-018-0320-3

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366:447–453. https://doi.org/10.1126/science.aax2342

Ochigame R (2019) The invention of "Ethical AI". https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/. Retrieved March 10, 2021

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown, New York

Overdorf R, Kulynych B, Balsa E, Troncoso C, Gürse S (2018) Questioning the assumptions behind fairness solutions. ArXiv:1811.11293. Retrieved March 11, 2021

Pariser E (2011) The filter bubble. Penguin, New York

Pasquale F (2015) The black box society: the secret algorithms that control money and information. Harvard University Press, Cambridge

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, pp 5684–5693.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH (2018) Ensuring fairness in machine learning to advance health equity. Ann Intern Med 169(12):866–872. https://doi.org/10.7326/M18-1990

Rawls J (1971) A theory of justice. Harvard University Press, Cambridge

Richardson R, Schultz J, Crawford K (2019) Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. N.Y.U. L. Review 94(192). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423. Retrieved March 10, 2021

Robbins S (2019) A misdirected principle with a catch: explicability for AI. Mind Mach 29(4):495–514. https://doi.org/10.1007/s11023-019-09509-3

Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. Knowl Eng Rev 29(5):582–638. https://doi.org/10.1017/S0269888913000039

Sandel M (1984) The procedural republic and the unencumbered self. Polit Theory 12: 81–96. http://www.jstor.org/stable/191382. Retrieved March 11, 2021

Sangiovanni A (2017) Humanity without dignity. Moral equality, respect, and human rights. Harvard University Press, Cambridge

Selbst AD, Boyd D, Friedler AS, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, 59–68. ACM Press. Atlanta, GA, USA: https://doi.org/10.1145/3287560.3287598.

Seng Ah Lee M, Floridi L (2020) Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Minds Mach. https://doi.org/10.1007/s11023-020-09529-4

Shah H (2018) Algorithmic accountability. Philos Trans R Soc Math Phys Eng Sci 376(2128):20170362. https://doi.org/10.1098/rsta.2017.0362

Shapiro S (2020) Algorithmic television in the age of large-scale customization. Televis New Med 21(6):658–663. https://doi.org/10.1177/1527476420919691

Shelby T (2016) Dark ghettos: injustice, dissent, and reform. Harvard University Press, Cambridge

Shin D, Park YJ (2019) Role of fairness, accountability, and transparency in algorithmic affordance. Comput Hum Behav 98:277–284. https://doi.org/10.1016/j.chb.2019.04.019

Simonite T (2020) Meet the secret algorithm that's keeping students out of college. Wired. https://www.wired.com/story/algorithm-set-students-grades-altered-futures/. Retrieved March 11, 2021

Scheffler S (2003) What is egalitarianism?. Philos Public Affairs 31(1): 5–39. http://www.jstor.org/stable/3558033. Retrieved March 11, 2021

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7

Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, Ho RCM (2019) Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med. https://doi.org/10.3390/jcm8030360

Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, Floridi L (2021) The ethics of algorithms: key problems and solutions. AI Soc. https://doi.org/10.1007/s00146-021-01154-8

Tufekci Z (2015) Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. J Telecommun High Technol Law 13(203). https://ctlj.colorado.edu/wp-content/uploads/2015/08/Tufekci-final.pdf. Retrieved March 11, 2021

Turner Lee N (2018) Detecting racial bias in algorithms and machine learning. J Inf Commun Ethics Soc 16(3):252–260. https://doi.org/10.1108/JICES-06-2018-0056

Umbrello S (2020) Imaginative value sensitive design: using moral imagination theory to inform responsible technology design. Sci Eng Ethics 26(2):575–595

Umbrello S, van de Poel I (2021) Mapping value sensitive design onto AI for social good principles. AI Ethics 1(3):1–14. https://doi.org/10.1007/s43681-021-00038-3

Van den Hoven J, Vermaas PE, van de Poel I (2015) Handbook of ethics, values, and technological design. Sources, theory, values and application domains. Springer. ISBN: 978-94-007-6969-4

Vyas DA, Eisenstein LG, Jones DS (2020) Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N Engl J Med 383(9):874–882. https://doi.org/10.1056/NEJMms2004740

Waldron J (2017) One another's equal. The basis of human equality. Harvard University Press, Cambridge

Williams B (1981) Persons, character and morality. Moral Luck: Philosophical papers 1973–1980. Cambridge University Press, Cambridge, pp 1–19

Wolff J (1998) Fairness respect, and the egalitarian ethos. Philos Public Affairs 27(2):97–122. https://doi.org/10.1111/j.1088-4963.1998.tb00063.x

Wolff J (2010) Fairness, respect, and the egalitarian "ethos" revisited. J Ethics 14(3/4):335–350

Wong P (2019) Democratizing algorithmic fairness. Philos Technol. https://doi.org/10.1007/s13347-019-00355-w

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2015) Fairness constraints: Mechanisms for fair classification. arXiv preprint arXiv:1507.05259.