Feminist Philosophy Quarterly

Volume 8 | Issue 3/4 Article 3

2022

Artificial Intelligence in a Structurally Unjust Society

Ting-An Lin
Stanford University
tinanlin@stanford.edu

Po-Hsuan Cameron Chen
cameron.ph.chen@gmail.com

Recommended Citation
Lin, Ting-An, and Po-Hsuan Cameron Chen. 2022. "Artificial Intelligence in a Structurally Unjust Society." *Feminist Philosophy Quarterly* 8 (3/4). Article 3.

Artificial Intelligence in a Structurally Unjust Society

Ting-An Lin and Po-Hsuan Cameron Chen

Abstract

Increasing concerns have been raised regarding artificial intelligence (AI) bias, and in response, efforts have been made to pursue AI fairness. In this paper, we argue that the idea of structural injustice serves as a helpful framework for clarifying the ethical concerns surrounding AI bias—including the nature of its moral problem and the responsibility for addressing it—and reconceptualizing the approach to pursuing AI fairness. Using AI in health care as a case study, we argue that AI bias is a form of structural injustice that exists when AI systems interact with other social factors to exacerbate existing social inequalities, making some groups of people more vulnerable to undeserved burdens while conferring unearned benefits to others. The goal of AI fairness, understood this way, is to pursue a more just social structure with the development and use of AI systems when appropriate. We further argue that all participating agents in the unjust social structure associated with AI bias bear a shared responsibility to join collective action with the goal of reforming the social structure, and we provide a list of practical recommendations for agents in various social positions to contribute to this collective action.

Keywords: artificial intelligence, AI bias, AI fairness, structural injustice, moral responsibility, health care

1. Introduction

Over the past decade, due to a surge in the use of artificial intelligence (AI) technology[1] (LeCun, Bengio, and Hinton 2015), increasing concerns have been raised regarding AI bias. Several studies have revealed that AI may reproduce existing social

[1] Since the term "artificial intelligence" was coined in the 1950s, a variety of methods have been developed to pursue the goal of building machines with some "intelligent capacities." The recent surge in AI has much to do with a type of technology named "machine learning," which trains a computer system with a huge quantity of data so that it will recognize some patterns in the existing datasets and then apply the learned patterns to make judgments regarding new cases. In this paper, we use "artificial intelligence (AI)" and "machine learning" interchangeably unless otherwise noted.

injustices, such as sexism and racism. For example, one AI recruiting tool downgrades resumes containing the keyword "women's," such as "women's college" and "women's chess club," resulting in a preference for male candidates (Dastin 2018); another algorithm used to judge risk for recidivism tends to falsely identify Black defendants as future criminals (Angwin et al. 2016); and the algorithms behind one global search engine tend to represent women of color with degrading stereotypes (Noble 2018).

In response to the growing concerns regarding AI bias, a burgeoning research paradigm under the names of *AI fairness*, *fairness in AI*, and *machine learning (ML) fairness* has been developed (Hajian and Domingo-Ferrer 2013; Kamiran, Calders, and Pechenizkiy 2013; Barocas and Selbst 2016; Lepri et al. 2018; Friedler et al. 2019). Additionally, many tech companies have launched relevant initiatives to pursue the value of AI fairness by developing fairer algorithms (Joyce et al. 2021). In the following text, we refer to this as the *dominant approach* to AI fairness. The dominant approach sees the problem of AI bias as being primarily located within algorithms, understands the goal of AI fairness as ensuring some parity of some statistical measures between different groups of people, and thus aims to fix the problem mainly through *debiasing* algorithms (Davis, Williams, and Yang 2021). Despite its popularity, critics have pointed out limitations of this dominant approach to AI fairness, including challenges of deciding between various statistical measures (Miconi 2017; Friedler, Scheidegger, and Venkatasubramanian 2016), concerns that AI systems are examined as entities isolated from the situated social contexts (Hoffmann 2019; Fazelpour and Lipton 2020; Le Bui and Noble 2020), and an overemphasis on technocentric responses (Fazelpour and Lipton 2020; Le Bui and Noble 2020). These reflections reveal a need for a more comprehensive moral framework to reconceptualize the ethical concerns and approaches involved in addressing AI bias and pursuing AI fairness. As suggested by Le Bui and Noble (2020, 163), "We are missing a moral framework" for thinking about ethics and justice in AI.

This paper argues that a conceptual framework named *structural injustice*, originally proposed by Iris Marion Young (2011), can serve as a helpful framework for clarifying the ethical concerns surrounding AI bias—including the nature of its moral problem and the responsibility for addressing it—and reconceptualizing the approach to pursuing AI fairness. In other words, this paper aims to conduct what Sally Haslanger (2000) calls an *ameliorative project* for the notions of AI bias and AI fairness by asking how these concepts should be used to better serve the purposes of using them.[2] We suggest understanding AI bias as a form of structural injustice that exists when AI systems interact with other social factors to exacerbate existing social inequalities, making some groups of people more vulnerable to undeserved burdens

[2] We thank Shen-yi Liao for this point.

while conferring unearned benefits to others. From this perspective, the notion of AI fairness should be about pursuing a more just social structure, potentially through the development and use of AI systems when appropriate. By situating AI systems within existing social structures, the structural-injustice perspective enables a more nuanced analysis of the contributing factors to AI bias and indicates potential directions for the pursuit of AI fairness. Drawing on Young's (2011) *social connection model* (SCM) of responsibility, we argue that a broader group of people beyond software engineers are responsible for pursuing AI fairness and should join collective action to shape the social structure. Overall, the structural-injustice framework helps clarify the moral problem of AI bias and offers new directions for more holistic responses to it.

This paper proceeds as follows. In section 2, we begin with a brief introduction to the idea of structural injustice. In section 3, we analyze the issue of AI bias through the lens of structural injustice. There, we use AI in health care as a case study and indicate several situations where interactions between AI development and existing social factors function to exacerbate health disparities. In section 4, we discuss how identifying AI bias as a case of structural injustice can help reconceptualize a *structural-injustice approach* to AI fairness. The last two sections consider moral responses to AI bias and the pursuit of AI fairness. In section 5, we discuss the insights provided by the SCM in terms of responsibility for AI bias. In section 6, we discuss the practical implications of the structural-injustice approach by presenting a list of recommendations for actions throughout the process of AI system development and noting points of interventions that agents in various social positions may take to approach the goal of AI fairness.

2. The Idea of Structural Injustice

The idea of structural injustice was proposed by Iris Marion Young to refer to a kind of moral wrong that cannot be reduced to individual wrongdoings or repressive policies but must be examined in terms of the influence of social structures. According to Young (2011, 52), structural injustice exists when "social processes put large groups of persons under systematic threat of domination or deprivation of the means to develop and exercise their capacities, at the same time that these processes enable others to dominate or to have a wide range of opportunities for developing and exercising capacities available to them." Structural injustice is present in various aspects of the modern world, including sexual and racial violence, homelessness, and labor injustice within global sweatshops.

Structural injustice concerns the *way* in which social structure restricts people's opportunities. Following Young, we can understand social structures as background conditions formed through complicated and dynamic interactions between *norms* (e.g., institutionalized laws and social norms), *schemas* (e.g., associated symbols, meanings, and values), and *distributions of resources* (e.g., public

transportation and health-care coverage).[3] By setting up the criteria of norms, associating meaning with different practices, and distributing resources to different groups of people, social structures affect the options available to people and thereby enable or constrain their actions. When a social structure unfairly restricts the options of certain groups of people and exposes them to the threat of undeserved burdens[4] while also conferring unearned privilege and power to other groups, this social structure is unjust and should be rectified.

Two notable features of structural injustice should be touched upon. First, structural injustice can result from interactions between morally permissive actions. While several actions may seem permissive when evaluated separately, they could still result in structural injustice when interacting with each other. Young's example of Sandy illustrates this point. A fictional character named Sandy is forced into homelessness due to the combined effect of many factors, including the cost of rental housing, the cost of a security deposit, the need to live in a certain location that she considers suitable for her children and her commute to work, and so on. In this case, Sandy's homelessness is not a direct result of deviant individual wrongdoings (e.g., mistreatments by landlords or real estate agents) or repressive policies (e.g., discriminatory policies blocking Blacks and Jews from owning or renting housing in certain neighborhoods). Many factors—such as the cost of housing, the location of mass transportation, and the availability of job openings—may not seem morally problematic when evaluated separately. However, when these multiple components interact in a complicated social process, the resulting social structure may assign certain groups of people to social positions that render them unjustly vulnerable to homelessness. The example of Sandy demonstrates that structural injustice is a moral wrong distinct from individual wrongdoings or repressive policies. Furthermore, this example emphasizes that solely evaluating each component of the social structure in isolation is insufficient to analyze the influence of the overall social structure. Instead, it is essential to take a broader perspective and evaluate the influence of the overall social process to effectively identify structural injustice.

Second, despite the "structural" aspect of the term "structural injustice," it is undeniable that individuals play a crucial role in this moral wrong. It is important to note that social structure exists only "in action" (Young 2011, 53) and is created and maintained as a cumulative result of individuals' actions. While most social structures

[3] The components of social structure listed here are largely drawn from Young (2011) and Haslanger (2016).

[4] While domination is emphasized in Young's original definition of structural injustice, we suggest that structural injustice could bring about other forms of undeserved burdens, such as exploitation and alienation. See McKeown (2016) for discussions on structural exploitation and Lu (2017) for discussions on structural alienation.

are not intentionally designed, individuals play critical roles in producing, sustaining, and shaping them. By following or obeying norms, embracing or challenging shared values, and distributing resources differently, individuals, through their everyday practices, may support or weaken the social structure. Therefore, the term "structural injustice" does not imply that individuals play no part in it. Instead, the structural-injustice framework offers a perspective that sees individuals not as discrete entities but as always connected to the broader social structure.

3. AI Bias as Structural Injustice: The Example of AI in Health Care

Having clarified the idea of structural injustice, we now argue that the issue of AI bias can be understood as an example of structural injustice. From this viewpoint, AI bias exists when AI systems interact with other social factors to exacerbate existing social inequalities, making some groups of people more vulnerable to undeserved burdens while conferring unearned benefits to others. This can occur when a system is developed and used in various social domains, such as hiring, parole decisions, health-care provision, and university admission. In this section, we use AI in health care as an illustrative example of this phenomenon.

Racial, gender, and class disparities in health care are well documented (Nelson 2002; Good et al. 2005; Manuel 2018). While AI has the potential to mitigate inequalities (I. Y. Chen, Joshi, and Ghassemi 2020), interactions between health-care AI development and other social factors often exacerbate existing health disparities. Below, we discuss several ways in which these interactions may occur throughout the process of developing health-care AI. Although this is not intended to be an exhaustive list, by indicating the various ways in which these interactions can occur, we hope to present a nuanced view of how AI bias may arise. These reflections also highlight ways in which the design of AI systems could potentially mitigate existing health disparities (to be discussed further in section 6.2).

3.1. Four Stages of AI Development

The developmental process of health-care AI roughly consists of four stages: *problem selection*, *data curation*, *algorithm development and validation*, and *deployment and monitoring* (P.-H. C. Chen, Liu, and Peng 2019).[5] Here, we use the

[5] It should be noted that the process of AI development described herein concerns one category of machine learning methods named *supervised learning*, in which both input and output data are provided for the algorithm to learn the relationship between them. While there are some other categories of machine learning methods, such as unsupervised learning and reinforcement learning, most AI systems in health care are currently developed based on supervised learning. Thus, this paper focuses on methods of supervised learning.

diagnosis of diabetic retinopathy from fundus images as an example to explain these four stages (Gulshan et al. 2016). Diabetic retinopathy is an eye complication that often occurs in diabetic patients and can cause blindness. For diabetic patients, diabetic retinopathy screening should be conducted annually. In this screening process, ophthalmologists evaluate fundus images and assign severity scores suggesting different downstream actions. For patients with low scores, no immediate intervention is needed aside from a yearly follow-up; for patients with high scores, active intervention and treatment are advisable.

The first step of developing a health-care AI system is to select a problem and formulate it into a machine learning task. In this case, the problem is formulated as the need for the algorithm to take fundus images, automatically classify these images into different severity scores following clinical guidelines, and output the classified scores.

The next step after problem selection is data curation. To develop the algorithm and evaluate its performance, two separate input–output paired datasets are needed: a *development set* for developing the algorithm and a *validation set* for evaluating the algorithm's performance (P.-H. C. Chen, Liu, and Peng 2019). In this case, the inputs are the fundus images, and the outputs are the diagnostic severity scores produced by the ophthalmologists reviewing the input images. The process of assigning desired output-prediction results to the input data (in this case, assigning severity scores based on fundus images) is called *labeling*. Typically, the labels of the validation set are obtained through the best available method for establishing the target condition, which is also referred to as a *reference standard* in the medical literature or a *ground truth* in the AI literature.

In the third stage—algorithm development and validation—engineers or scientists design algorithms and train these algorithms based on the development set. Once algorithm development is complete, the final algorithm is applied to the validation set, and the predictions made by the algorithm are compared against the labels of the validation set to measure the algorithm's performance. The results represent the expected performance of the algorithm for an unseen dataset with a similar data distribution to that of the validation set.

Lastly, in the deployment and monitoring stage, the finalized algorithm is used in real-world workflows subject to potential regulatory oversight. For example, after a successful prospective pivotal trial (Abràmoff et al. 2018), the US Food and Drug Administration authorized an automated tool for diabetic retinopathy screening without the need for a clinician to interpret the results (FDA 2018). Deployment in real-world scenarios requires integration of the algorithm into the clinical workflow, ensuring that any personnel who may use the algorithm are well trained and that the algorithm applies to the intended population. To ensure consistent algorithm performance, postdeployment monitoring is needed.

3.2. How Health Disparities May Be Exacerbated throughout the AI-Development Process

Throughout the four stages of health-care AI development, there are numerous ways in which AI systems may interact with other social factors and intensify existing health disparities.

First, there is the problem-selection stage. As previously noted by feminist scholars, the selection of a health-care problem to address can be influenced by multiple factors, such as funding availability, data availability, and the interests of the decision-makers.[6] Many studies have revealed that funding allocation for health-care research is strongly imbalanced across different populations. As suggested by the *10/90 gap*,[7] funding for health-care research is disproportionately spent on problems that affect people in higher-income countries, whereas poverty-related diseases in the Global South (e.g., malaria) are relatively underfunded (Vidyasagar 2006). In the US, cystic fibrosis (a disease that mostly affects White populations) receives 3.4-fold greater funding per affected individual than sickle cell anemia (which mostly affects Black populations), despite the fact that both genetic disorders are of similar severity (Farooq and Strouse 2018).

Funding availability is not the only resource restriction that influences problem selection. Given the need for large datasets for AI development, the availability of digitized data—which is greatly influenced by existing health inequities—affects what problems can be addressed. For example, there is a discrepancy in the availability of electronic health records (EHRs) between high-income and low-income countries. As a result, the selected problems are more likely to be those that occur in high-income countries with EHRs. The interests of the people making problem-selection decisions also influence which problems are selected (I. Y. Chen et al. 2021, 11). The term *gender research gap* refers to the lack of research on women's health conditions and is related to the gender imbalance in the scientific workforce (Fisk and Atun 2009). A lack of diversity in the workforce making decisions regarding AI development may lead to biased attention in problem selection.

After a problem is selected, researchers move on to the second stage—data curation—in which two types of datasets (development and validation sets) are

[6] For more detailed discussions regarding the influence of values in the construction of scientific knowledge, see Longino (1990), Anderson (1995), and Wylie and Nelson (2007). Thanks to an anonymous reviewer for noting this connection and providing these references.

[7] The 10/90 gap refers to the phenomenon of disparity in health research spending between diseases that affect rich and poor countries. As Vidyasagar (2006, 55) puts it, "less than 10% of global funding for research is spent on disease that afflict more than 90% of the world's population."

developed. There are at least three different ways in which data may contribute to AI bias: (1) when the data used for training is insufficiently representative, (2) when existing biases are encoded in the data, and (3) when the reference-standard selection introduces bias.[8] The first form of data-related bias occurs when the data used for training are insufficiently representative, especially for marginalized groups. For example, in order to train an AI system for good performance in aiding health-care decisions, it is critical to include relevant data on health conditions. However, existing health data are often skewed toward more privileged populations. For example, many widely used open datasets are US- and Europe-centric and thus lack the geodiversity needed to develop AI systems for other parts of the world (Shankar et al. 2017). In fact, most of the medical data used to train AI algorithms are derived from three states in the US: California, Massachusetts, and New York (Kaushal, Altman, and Langlotz 2020). In the US, EHR datasets include largely White populations with private insurance and much fewer uninsured Black and Hispanic patients (Hing and Burt 2009). In the development of AI systems for skin cancer diagnosis, one of the largest databases used for model development collects pigmented lesions disproportionately from light-skinned populations (Adamson and Smith 2018). Such skewed data distributions produce AI systems with better performance for light-skinned individuals than for people of color.

Data can also become a source of bias when it incorporates existing systemic biases against marginalized groups. For example, studies have shown that many clinical notes incorporate the biases of medical practitioners against marginalized patients, such as African American and LGBT groups (Geiger 2003; Fallin-Bennett 2015). When these clinical notes are used to develop AI systems, the resulting systems will tend to replicate existing biases.

Finally, data-related bias may occur through the selection of reference standards. As previously mentioned, in algorithm development and validation, the reference standard serves as the ground truth for evaluating the algorithm's performance. The selection of reference standards often involves a subjective judgment, and the designers' perceptions of what should serve as the ground truth may be influenced by the existing social structure. For example, when training AI algorithms for disease diagnosis, diagnosis by a doctor is often used as a reference standard, under the implicit assumption that a doctor's diagnosis reflects objective facts about the severity of a certain disease (Pierson et al. 2021). However, doctors with different levels of training or from different countries may be more or less familiar with patients from certain demographic backgrounds, making their diagnoses more or less accurate for certain populations. Without taking these factors into

[8] Thanks to an anonymous reviewer for helping us clarify the three different ways in which data-related bias can occur.

account, the selection of reference standards could replicate existing health inequities. One study revealed that a certain algorithm widely used to evaluate health needs tends to assign lower scores to Black patients than to White patients (Obermeyer et al. 2019). This largely stems from the fact that the designer used health-care costs rather than illness as the reference standard for deciding health needs. Multiple social factors—including socioeconomic status, racial discrimination, and reduced trust in the health-care system—have contributed to the relatively low health-care costs spent on Black patients. As a result, the choice of health-care costs as a proxy for health needs fails to genuinely reflect health needs and intensifies racial disparities.

Although the third stage of algorithm development and validation may sound purely mathematical, it also involves decision-making influenced by existing social factors. It is often suggested that AI merely reflects the data distribution, as suggested by the slogan "Garbage in, garbage out." By focusing on the problem with data, the discussion often implicitly portrays the algorithms as objective in the sense that they are not the root of biased outcomes (Smith 2018). However, the design of an algorithm can amplify the disparate impact of social factors across different populations (Hooker 2021) and thus has the potential to amplify health disparities. For example, in deciding what kind of information should be used for algorithm training, the problem of whether and how demographic information should be incorporated is a complicated issue that poses serious challenges. In some cases, the incorporation of demographic information could lead the algorithm to exploit such information and thus replicate existing biases embedded in the data when making predictions. For example, vaginal birth after cesarean (VBAC) scores are assigned by an algorithm designed to predict the success rate of childbirth with prior cesarean section. Based on observational studies that have revealed a "correlation between racial identity and the success rate," it can be concluded that VBAC scores "explicitly include a race component as an input" (I. Y. Chen et al. 2021, 11) and assign lower VBAC scores to Black and Hispanic women compared with White populations (Vyas, Eisenstein, and Jones 2020). While the underlying cause of this correlation between racial identity and success rate is unclear, this association may be due to the unequal distribution of medical resources among different racial groups. By directly using VBAC scores as part of its training information, the developed algorithm would replicate the pattern of racial disparity, predict a lower success rate of trial of childbirth labor for Black and Hispanic women, and thus further exacerbate existing health inequalities.

While the example above shows that incorporating demographic information could worsen existing inequalities, it does not imply a necessary link between the inclusion of demographic information and the replication of existing inequalities. On the one hand, even if demographic information is not directly incorporated,

confounders between the data and sensitive demographic information (e.g., ZIP code is often associated with the residents' race/ethnicity demographics) may persist and thus function to replicate existing patterns (Rodriguez et al. 2007; Johnson 2021). On the other hand, some approaches have explicitly incorporated demographic information to ensure that the developed algorithms do not utilize this information in making predictions (Zhao, Adeli, and Pohl 2020). These observations indicate that deciding whether and how to incorporate sensitive information requires careful consideration regarding the nature of the information.

In validating an algorithm's performance, the way in which the performance is measured and defined is another area involving normative judgment. The example of the COMPAS-ProPublica debate highlights the importance of this factor. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a recidivism prediction system used by US courts to analyze the likelihood that a defendant will commit new crimes in the future. In 2016, the website ProPublica investigated the algorithm for this system and argued that it is racially biased because Black defendants are "twice as likely as whites to be labeled a higher risk but not actually re-offend" (Angwin et al. 2016). Northpointe, the company that developed COMPAS, responded to this criticism and argued that the model made similar rates of mistakes between Black and White defendants. As noted by Hellman (2020, 816), "The controversy focused on the manner in which such similarity is assessed." Northpointe focused on the predictability between the assigned score and the likelihood of recidivism, whereas ProPublica pointed out differences in false positive and false negative rates between Blacks and Whites. The lesson that can be derived from this debate is not an easy answer to the question of which measure is correct; rather, it highlights the complicated nature of making such a decision.

Finally, after an algorithm has been developed and validated, its deployment for real-world use can have a compounding effect on health disparities. For example, consistency between the population for which an algorithm is deployed and the originally intended use is essential to the algorithm's performance (Kelly et al. 2019). However, for some of the reasons mentioned above, there is a tendency to develop algorithms intended for use with widespread health conditions in privileged groups. As a result, it is easier to ensure consistency in these contexts than in other health conditions or populations.

Even if an algorithm were developed to improve health-care conditions for diverse groups of people—for example, to aid doctors in the diagnosis and treatment of cancer—this does not mean that different groups of people would benefit similarly from this algorithm. Rather, under current social processes, access to the potential benefits of this new technology could be influenced by existing resource distributions. Hospitals located in wealthy areas of high-income countries may have more resources to purchase and incorporate the algorithm into their services. Furthermore, in

hospitals with more resources, AI systems may play an assisting role for doctors in the process of medical diagnosis; in contrast, in places with more restricted medical resources, there might be a need to rely more heavily on AI systems to make diagnoses despite the quality of their performance. Overall, disparate access to and use of AI algorithms could compound with and intensify existing disparities in health resources.

In summary, while the analysis above is not intended to be an exhaustive list, it serves to reveal how AI development interacts with various social factors and may function to exacerbate existing health disparities. From a structural-injustice perspective, this analysis provides a nuanced view of how AI bias may occur by analyzing it against the situated social context, thus drawing attention to the interactions between AI systems and other social factors. From this perspective, AI bias is better understood not as a feature of AI algorithms but as the way in which AI systems are embedded in and reinforce the unjust social structure, in which certain groups of people are unfairly exposed to undeserved burdens. The influences between AI systems and existing social injustices are bidirectional. In the case of AI in health care, several structural gender, race, and class inequalities influence the development and application of AI systems, exacerbating existing health disparities. The deployment of these AI systems often strengthens and materializes oppressions (Liao and Carbonell 2022) along the axes of gender, race, class, and so on. In this way, AI systems and other forms of social injustice reinforce each other and sustain oppressive social structures.

4. Toward a Structural-Injustice Approach to AI Fairness

If AI bias is recognized as a case of structural injustice, then we should also reconceptualize the pursuit of AI fairness. Below, we discuss several lessons regarding the goal and methods of AI fairness from the structural-injustice perspective and contrast this perspective with the dominant approach to AI fairness.

4.1. AI Fairness Is More Than Ensuring Statistical Parity

First, let us reconsider the goal of AI fairness. The dominant approach portrays AI bias as a problem that is mainly located within the AI system and thus focuses on fixing problematic algorithms. This approach tends to examine the outcomes of AI systems as entities isolated from background conditions and sees the goal of fairness as ensuring the parity of some statistical measures (Davis, Williams, and Yang 2021). Critics have pointed out that overemphasis on the discrete analysis of “bad algorithms” pays insufficient attention to the related social and cultural contexts in which these AI systems are situated (Hoffmann 2019) and fails to capture the power dynamics that AI systems are embedded in and interact with (Le Bui and Noble 2020). In contrast, the identification of AI bias as a case of structural injustice suggests that the moral

problem is not merely located within the AI systems but also results from interactions of the overall social structure, of which AI is one component. By situating AI systems against existing social contexts, the structural-injustice perspective understands the goal of AI fairness differently from the dominant perspective.

First, by paying attention to the overall impact of interactions between AI systems and many other social factors, the structural-injustice perspective highlights that ensuring the statistical parity of a model's performance may be insufficient to achieve AI fairness. Given the existing social inequalities, AI systems that "perform similarly well" for different groups may function to reproduce the unjust social structure. Instead, the more appropriate goal of AI fairness should be understood as the pursuit of a more just social structure—or a social structure with fewer unjustifiable power imbalances—through the development and use of AI systems when appropriate. Alternatively, considering the ultimate goal of structural justice, this perspective may instead suggest that we should not develop or use AI systems in certain domains.

The analysis of AI bias through the structural-injustice framework draws attention to aspects of justice that go beyond resource distribution. Current discussions surrounding AI bias often emphasize the unequal distribution of goods (Hoffmann 2019), such as job opportunities, admissions, and medical resources. It is often suggested that the main problem of AI bias is that it outputs an unequal distribution of resources for different groups of people. Based on this understanding, the aim of fixing AI bias is to ensure the parity of distribution across different groups of people. For example, Heidari and colleagues suggested that the notion of fairness should be understood as "equality of opportunity" (Heidari et al. 2019),[9] and Rajkomar and colleagues suggested that AI systems should be made to implement "principles of distributive justice" in terms of ensuring "equality in patient outcomes, performance, and resource allocation" (Rajkomar et al. 2018, 886).

Given that the moral problem of structural injustice lies in the imbalance of power relations between groups of people and their situated social positions, structural analysis calls attention to the variety of components that shape the current power structure. Instead of merely asking whether the performance of an AI system, evaluated in isolation, produces fair distributions of goods across different groups of people, structural reflections suggest that the following questions should be raised: How does the AI system, as part of the social structure, shape the existing power structure between different groups of people? Does it function to reduce, reproduce, or further exacerbate existing power imbalances? This perspective naturally broadens its attention beyond resource distribution to also consider issues such as recognition and representation, which also contribute to the hierarchies of social status (Fraser

[9] We thank an anonymous reviewer for contributing this reference.

and Honneth 2003). When a comprehensive self-tracking app failed to include a menstruation tracking function (Eveleth 2014), it conveyed the message that women's health experiences were not important. When a global search engine associates Black women with degrading stereotypes (Noble 2018), they suffer from representational harm. In these cases, even though the harm is not mediated through unequal resource distribution, the use of these systems reinforces unjust power imbalances and thus should be addressed according to the structural-injustice approach to AI fairness.

Furthermore, by paying attention to unfair power imbalances, the structural-injustice approach expands its focus beyond the output of AI performance to also consider the power relations embedded in the process of AI development. Accordingly, this approach raises the following questions: How are decisions regarding the development of new technological systems made? Who is given a say in the process? How does the decision-making process function to support or shape existing power relations? The structural-injustice approach promotes closer scrutiny of the development and decision-making process to avoid the replication of unjustified power relations through this process.

4.2. AI Fairness Cannot Be Achieved through Debiasing Algorithms

Clarifying the moral nature of AI bias not only helps in identifying related problems that should be addressed, but it also helps in creating better approaches to address AI bias and pursue the goal of fairness. The dominant approach sees the problem of AI bias as primarily computational or mathematical and thus promotes solving AI bias mainly through the technological advancement of debiasing algorithms (Davis, Williams, and Yang 2021). Concerns have been raised regarding this overall trend of turning to software engineers and technology corporations as the go-to people for solving this issue. For example, (Le Bui and Noble 2020, 177) have expressed concern that the overemphasis on technological solutions could lead us to become *techwashed*—that is, overly rely on techno-centric responses.

Although technological tools should certainly be part of the response to AI bias, they are not a full solution. For one, the task of deciding which statistical measures should be used to build debiasing or fairer algorithms is an issue that goes beyond computations (recall the COMPAS-ProPublica debate mentioned in section 3.2). While a growing number of matrices of fairness have been proposed, it is mathematically impossible for an algorithm to satisfy all these different matrices simultaneously (Friedler, Scheidegger, and Venkatasubramanian 2016; Miconi 2017). Making the trade-off between these different statistical metrics requires insights from both ethical and social perspectives.

Regarding the claim that it is impossible to satisfy all these statistical measures, Hedden (2021) recently argued that this does not necessarily entail making a trade-

off, as most of these statistical criteria are not necessary conditions for fairness. The implication of Hedden's analysis is not that we can easily pursue the goal of fairness by focusing on one or a few statistical measures of an algorithm's performance. Rather, it points out that the source of bias and unfairness may not lie in the algorithm itself but "could instead lie elsewhere: with the background conditions of society, with the way decisions are made on the basis of its predictions, and/or with various side effects of the use of that algorithm, such as the exacerbation of harmful stereotypes" (Hedden 2012, 227). This point resonates well with the structural-injustice perspective's emphasis on the overall influence of social structure, of which the algorithm is merely one component.

Given that structural injustice has resulted from interactions between AI algorithms and other social factors, the problem of AI bias cannot be fixed solely by designing algorithms with statistical parity. Instead, more diverse approaches examining social, political, and institutional structures should also be developed to intervene in the power imbalances surrounding AI development and use. From this perspective, deferring to tech developers and companies for solutions to AI bias is inadequate and risks conferring unearned power to a small group of people. To move forward, the following questions should be reconsidered: Who should be responsible for addressing AI bias and pursuing fairness? Why should they be responsible? What does being responsible for this issue entail? The following two sections discuss these issues in greater detail.

In summation, while the dominant approach tends to see the goal of AI fairness as ensuring the statistical parity of AI model performance, the structural-injustice approach emphasizes the goal of pursuing a more just social structure, potentially including the development and use of AI systems when appropriate. By recognizing this goal, the structural-injustice approach suggests that AI fairness cannot be achieved by mainly relying on computational tools or deferring to software engineers for solutions. Rather, this approach highlights the need to address other factors of the social structure and appeal to more diverse approaches to intervene in the power imbalances surrounding AI development and use.

5. Responsibility for AI Fairness: Insights from the Social Connection Model

We have argued that structural injustice is a helpful framework for analyzing the ethical problem of AI bias and reconceptualizing the idea of AI fairness. Now, we will discuss what this reconceptualization implies regarding the corresponding responsibility for AI fairness. This section will begin by highlighting the challenges of responsibility attribution based on the conventional *liability model.* Due to the features of AI development, the conditions for assigning moral responsibility are somewhat difficult to meet. While we should not necessarily give up the task of responsibility attribution, the recognition of AI bias as a case of structural injustice

suggests that we should consider appealing to other moral grounds in responsibility attribution. Thus, we turn to Young's SCM of responsibility and discuss how it advances existing discussions on responsibility for AI bias. Drawing from the SCM, we suggest that all agents participating in the unjust social structure associated with AI bias bear a *shared responsibility* for this structure and should join collective action for its reform.

5.1. Challenges Encountered by the Liability Model

Two conditions are typically required to assign moral responsibility based on the liability model: the *control condition* and the *knowledge condition* (Coeckelbergh 2020). The control condition suggests that one is responsible for an action if one is the agent of the action, has caused the action, or has a sufficient degree of control over the action. The knowledge condition suggests that one is responsible for an action if one knows what one is doing in terms of awareness of the action, the moral significance of the action, the consequences of the action, and the alternatives (Rudy-Hiller 2018). If an agent's action satisfied both conditions, this agent would be held responsible in the liability sense and would typically be subject to blame and punishment.

In considering the impact of AI systems and attributing responsibility to the designers, these two conditions are somewhat difficult to meet. The control condition encounters the problem of *many hands* (van de Poel et al. 2012), meaning that too many agents are involved in the development of AI systems. Additionally, the control condition encounters the problem of *many things*, meaning that many technologies are involved in an AI system's development (Coeckelbergh 2020). As illustrated in section 3, the complicated interactions between factors throughout the process of AI development and deployment make it difficult to establish causal chains for the control condition. Even if some of the causal chains could be traced, it could be the case that none of the agents involved in development met the criteria of having a sufficient degree of control over the final impact. Moreover, the so-called *black box problem* highlights that AI algorithms are often neither fully transparent nor completely explainable, even to the engineers building them. This feature of AI algorithms makes it difficult for the knowledge condition to be satisfied and thus constitutes a *responsibility gap* (Matthias 2004).

These challenges do not mean that the idea of responsibility attribution should be abandoned. In fact, some have proposed counterarguments to the claim of a responsibility gap in AI development, many appealing to plural notions of responsibility that include but also extend beyond accountability (Köhler, Roughley, and Sauer 2017; Coeckelbergh 2020; Tigard 2021). Along these lines, we argue that the structural nature of AI bias warrants consideration of an additional notion of responsibility based on social connections.

5.2. Responsibility for AI Fairness Based on Social Connections

In response to discussions surrounding the activist movement against garment sweatshops, Young (2006) argued that a common conception of responsibility (the liability model) was suitable for assigning responsibility for individual wrongdoings but inappropriate for assigning responsibility for structural injustices. Several central features of structural injustice make it difficult to satisfy the control and knowledge conditions, thus rendering the application of the liability model unsuitable. First, given that structural injustice is a collective result, it is nearly impossible to trace the exact causal chain of one individual's action to the resulting injustice in a way that demonstrates that this action is the sufficient cause of the injustice. Furthermore, agents typically contribute to the reproduction of the social structure through their everyday practices without awareness of the consequences of their contributions. However, failure to meet the control and knowledge conditions does not suggest that an individual has no involvement in the resulting social structure. Instead, as noted in section 2, social structure only exists as a result of participation by various individuals. Through participation in the social process, individuals' everyday practices—whether intentional or not—play a critical role in reproducing and sustaining the social structure. Such contributions to the social structure constitute a new moral ground for attributing responsibility.

To accommodate the features of structural injustice, Young proposed a new model of responsibility attribution called the SCM.[10] The SCM suggests that agents who are "connected" with the social structure, or whose actions contribute to the reproduction of the social structure (McKeown 2018), bear some sort of responsibility (which Young called *political responsibility*) for structural injustice. Young suggested that political responsibility does not concern fault-finding and blame but mainly concerns more forward-looking goals, such as reformation of the social structure. Furthermore, given that no individual has the power to reform the social structure alone, political responsibility is a shared responsibility, meaning that it can only be discharged through collective action. People who bear this shared responsibility should attempt to coordinate, initiate, or participate in collective action to reform the social structure. The goal of this collective action is to ensure that the power imbalances between different social positions are removed, such that no one will be

[10] Young emphasized that she did not intend to replace the liability model with the SCM but rather intended for the SCM to serve as a complementary conception of responsibility that could be applied in contexts for which the liability model is not appropriate. In other words, according to Young, the liability model is appropriate for assigning responsibility for individual wrongdoings, whereas the SCM is appropriate for assigning responsibility for structural injustice.

put into social positions where they would be vulnerable to undeserved harm, such as domination, violence, and exploitation.[11]

In the case of AI bias, like many other structural injustices, the everyday practices of participants in the social structure often fuel its reproduction. Thus, according to the SCM, nearly all participants in the social structure are responsible for AI bias. In other words, the bearers of political responsibility for addressing AI bias and pursuing fairness include a vast group of people, from the CEOs of tech companies and the engineers involved in the development process to the government and ordinary people. It is true that many agents who contribute to the AI-development process may have little control over the actions available to them. It is also true that many of these agents do not intend to make a wrongful impact. Nevertheless, by using services supported by AI systems, by passively allowing decisions to be made by others, and by not intervening in the current process of AI development, our actions and inactions contribute to the reproduction of the social structure that results in structural injustice. In this way, participants in the social structure associated with AI bias can be described as *structurally complicit* (Aragon and Jaggar 2018) and should thus be held politically responsible for the unjust results.

By providing new grounds for responsibility attribution, the SCM avoids the challenges that confront the liability model and further implies that a broader group of people bears responsibility for addressing AI bias and pursuing fairness. The idea that nearly everyone bears responsibility for AI bias has been proposed recently but typically for different reasons from those proposed by the SCM. For example, Zimmermann, Di Rosa, and Kim (2020) suggested that "algorithmic bias is not a purely technical problem for researchers and tech practitioners" but a "moral and political problem in which all of us—as democratic citizens—have a stake." Similarly, Wong and Simon (2020) suggested that "it is essential for the ethics of AI to include various stakeholders, e.g., policy-makers, company leaders, designers, engineers, users, non-users, and the general public, in the ethical reflection of autonomous AI."

While we agree with this reasoning regarding democratic citizens and stakeholders, we wish to add that the reasoning underlying the SCM emphasizes that all individuals are agents with the capacity to shape the social structure. Although no individual can change the social structure independently, by working together with

[11] While Young's SCM is quite influential in discussions on structural injustice, this model and the corresponding notion of political responsibility have also received many critiques. For example, see Nussbaum (2009) for critiques of Young's claim that political responsibility does not concern blame; see Gunnemyr (2020) for critiques regarding Young's idea of connections; and see Schwenkenbecher (2021) for critiques regarding the lack of specificity on shared responsibility. Lin (forthcoming) also points out some theoretical gaps of Young's theory that require further explorations.

others, one may have the power to decide whether to develop, why to develop, and what to develop. From this perspective, ordinary individuals are not merely "patients" of responsibility who are affected by actions and decisions made by others and thus "may demand reasons for actions and decisions made by using AI," as suggested by Coeckelbergh (2020). Rather, ordinary individuals are also agents of responsibility who can act and make decisions regarding the development and use of AI.

6. Practical Implications for Measures to Pursue AI Fairness

The claim that nearly everyone bears responsibility for addressing AI bias and pursuing AI fairness may seem somewhat vague in terms of practical actions. One may ask, "So, what should I do? How can I contribute to this goal?" While case-by-case examinations are required to develop more detailed guidance, in this final section, we aim to provide a theoretical framework to aid in the direction of pursuing AI fairness.

6.1. A Division of Labor to Support Collective Action

According to the SCM, political responsibility is a shared responsibility that can only be discharged through collective action. Thus, bearing responsibility for AI bias implies joining in collective action to reform the social structure and thus prevent it from sustaining or exacerbating imbalanced power relations. Two features of this claim should be noted. First, the content of political responsibility is rather open in the sense that it does not directly specify the required actions. Instead, political responsibility identifies some ends (in this case, shaping the social structure into a less unjust form) for agents to use their own discretion in deciding their actions. Second, although political responsibility is shared among all participants in the social structure, the concept does not imply that responsibility is shared equally, nor does it imply that similar actions should be made by all responsible agents. Instead, one's situated social position is a crucial factor to consider.

An agent's situated social position greatly influences the power and resources available to the agent and thus has a substantive impact on the forms of action the agent can take to participate in a collective movement. Several proposals have been raised to help agents determine their relations to different kinds of structural injustice and decide what actions they should take. Young (2011) suggested four parameters of reasoning: power, privilege, interest, and collective ability. Agents with *power* are those with a great capacity to shape the social structure and should use this capacity accordingly; in the case of AI bias, leaders of tech companies and governments fall into this category. Agents with *privilege* are those who benefit from the current social structure, and these individuals should use the abundance of resources available to them to change their everyday practices and support the movement toward a more just social structure. Agents who are oppressed have some special *interest* in reshaping the social structure, and Young suggested that these individuals should also

contribute to the collective action such as by sharing their situated knowledge and experience to guide how the social structure should be reshaped. Lastly, agents with *collective ability* are those who belong to a group that can influence the process of reshaping the social structure and should direct their resources accordingly.

Young's four parameters of reasoning are not the only model we can use for reference. For example, Zheng (2018) provides the alternative *role-ideal model*, which suggests that individuals can better decide on the actions they should take by contemplating their social roles (e.g., as teachers, parents, employees, or citizens). According to this perspective, scholars should conduct in-depth examinations of the interactions between AI systems and existing social factors to propose diverse intervening approaches; together with the CEOs of tech companies, engineers and tech developers should try to develop AI systems that can help mitigate existing power imbalances; citizens of democratic societies should raise concerns regarding the design, development, and deployment of AI as parts of democratic agendas; and governments and lawmakers should work to provide suitable institutional designs that are conducive to these practices.

6.2. Recommendations for the Pursuit of AI Fairness: AI in Health Care as an Example

As noted by Wawira Gichoya and colleagues, while a few guidelines for designing AI systems have recently been proposed, there is still a lack of operationalizing recommendations for the pursuit of increased AI fairness (Wawira Gichoya et al. 2021). In response, they propose a handful of recommendations, including suggestions such as "engage members of the public in the process of determining acceptable standards of fairness" and "collect necessary data on vulnerable protected groups" (Wawira Gichoya et al. 2021, 2). Although we generally agree with their recommendations, they provide little reasoning in support of these recommendations. To address this gap and advance the discussion on this issue, we aim to provide a more comprehensive (but by no means exhaustive) list of recommendations for the pursuit of AI fairness, derived from structural-injustice analysis. Table 1 is an overview of these recommendations, organized along the four stages of AI development, as discussed in section 3. Below, we discuss these recommendations and provide examples of how different moral agents may help in achieving the goal of AI fairness.

Stages	Recommended Actions
Stage 1. Problem Selection	● Assess relevant social contexts and existing social inequalities to identify the potential risks and benefits of developing AI systems. ● Engage members of diverse groups (especially marginalized groups) in deciding on problems to address. ● Assess overall resource distributions to avoid enlarging resource gaps between groups.
Stage 2. Data Curation	● When choosing datasets, evaluate them to ensure that their data distributions are representational with no embedded biases against marginalized groups. ● Evaluate the selected reference standard to avoid replicating existing inequality.
Stage 3. Model Development and Validation	● When deciding what information should be used for algorithm training and what measure should be used for validation, critically analyze the associated social factors. ● Engage diverse groups of stakeholders in the process of determining acceptable measures for evaluating model performance.
Stage 4. Model Deployment and Monitoring	● Before deployment, assess consistency between the originally intended use and the population for which the algorithm will be deployed. ● After deployment, constantly evaluate the model's real-world impacts and make adjustments accordingly.

Table 1. List of Recommendations for the Pursuit of AI Fairness

For the first stage—problem selection—we provide three recommendations that emphasize paying attention to decisions that are rarely made with transparency. First, we suggest that in deciding whether to develop an AI system for a certain use, efforts should be made to assess relevant social contexts and existing social inequalities in order to identify the potential risks and benefits that such a system may bring about. Empirical work in the social sciences and other areas could play a crucial role in these efforts (Joyce et al. 2021), for example, by noting the unexplained

pain levels of underserved populations and thus motivating engineers to design AI systems that may better assist the diagnosis process in reducing health disparities (Pierson et al. 2021). Second, attention should be paid to power dynamics throughout the decision-making process. This process should engage members of diverse groups, especially marginalized groups, in identifying and deciding on problems to address. For example, the tech industry could pursue this goal by striving toward a more diverse and demographically representative composition. As feminist scholars have long argued, oppressed individuals, through being situated in marginalized positions, have distinct access to knowledge regarding oppression (Collins 1990; hooks 1990). Thus, incorporating the viewpoints of oppressed individuals could make the decision-making process more attentive to existing social inequalities. Furthermore, the act of including marginalized people in this process recognizes their agency and can thus function as an empowering experience. Third, efforts should be made at the institutional level to ensure that the resource distribution is not overly skewed toward privileged groups, as in the 10/90 gap in health-care research funding.

For the second stage—data curation—we propose two recommendations. First, in choosing and curating datasets, evaluations are required to ensure that their data distributions are representational with no biases against marginalized groups embedded in the data. Some efforts should be made by designers (e.g., to choose more representative datasets for training systems), but this goal can only be achieved if efforts are made at the institutional level to collect more representative datasets and make them available. Second, reflection is required in choosing a reference standard to avoid replicating the associated injustices against marginalized people.

A recent study by Pierson and colleagues (Pierson et al. 2021) serves as a good example of how paying proper attention to these two aspects of developing AI systems has the potential to help mitigate existing health disparities. It is well documented that underserved populations (e.g., people of color, lower-income patients, and less educated patients) experience higher levels of pain due to diseases such as osteoarthritis of the knee (Poleshuck and Green 2008; Allen et al. 2009; Eberly et al. 2018). To inquire about the causes of such pain disparities, Pierson and colleagues developed AI systems that take knee X-ray images as input and predict pain scores as output (Pierson et al. 2021). Being aware of the potential influence of existing human biases on the development of standard clinical guidelines, they intentionally chose not to use physicians' diagnoses as the reference standards for their algorithm, as this could replicate the existing biases embedded in established medical knowledge. Instead, they chose to train the algorithm based on a dataset of knee X-ray images with high racial and socioeconomic diversity, using patients' reported pain scores as labels. The resulting AI system predicted pain scores with a much lower percentage of unexplained pain disparities than the diagnoses made by physicians following standard clinical guidelines. This result revealed that information

regarding the severity of osteoarthritis was, in fact, available in the X-ray images, even though the standard clinical guidelines developed decades prior (based mostly on the diagnosis of White British populations) failed to capture this information. This suggests that much of the unaccounted pain of underserved populations is related to the racial and socioeconomic biases embedded in the standard clinical guidelines, which fail to capture some physical causes of pain in underserved groups and thus result in misdiagnosis of the severity of their osteoarthritis. If integrated into clinical practice, the AI system developed by Pierson and colleagues could potentially ensure that the reported pain levels of underserved populations are taken more seriously and that the severity of their osteoarthritis is measured more accurately (Pierson et al. 2021).

In the third stage—algorithm development and validation—it is crucial to recognize that choices about what information should be used and what criteria should be adopted are value-laden. First, in deciding what kind of information should be used to train an algorithm, attention should be paid to the critical assessment of potentially associated social factors (e.g., connections between ZIP codes and racial/ethnic demographic distributions) to avoid replicating existing inequality. Furthermore, in the process of algorithm validation, extra consideration should be paid to social background in determining which criteria should be used as measures (a lesson learned from the COMPAS-ProPublica debate).

While some aspects of decision-making require special expertise and thus are confined to the "experts," we suggest that at least some part of the decision-making process should incorporate members of the general public, who are highly likely to be impacted by AI systems. Some recent attempts to accomplish this have been made through the establishment of a form of political institution called *deliberative minipublics*. In general terms, deliberative minipublics aim to form microcosms of the public (i.e., convene an assembly of people who demographically represent the public) and provide these representatives with sufficient information and time to deliberate on issues of public concern and obtain results to inform the broader public for relevant discussions (Dahl 1989; Escobar and Elstub 2017). Over the past few decades, hundreds of minipublics have been established to address various social issues, recently including the public's perspectives regarding the development and use of AI. In 2019, the National Institute for Health Research in the UK organized two Citizens' Juries on Artificial Intelligence, in which groups of randomly selected citizens convened for several days to learn, deliberate, and produce a final report regarding the use of AI and the trade-off between the accuracy and explainability of AI systems (van der Veer et al. 2021). The results revealed that people's preferences regarding this trade-off differed across different contexts.

For the final stage—algorithm deployment and monitoring—we propose two recommendations. First, before the deployment of an AI system, consistency should

be assessed between the originally intended use and the population for which the system will be deployed, as major inconsistencies between these factors generally worsen the performance of an AI system. Second, after deployment, efforts should be made to constantly evaluate the system's real-world impacts and make any necessary adjustments. One important lesson that can be drawn from structural-injustice analysis is that, since unjust results are formed through complicated dynamics between many social factors, new forms of injustice could still occur even if careful examinations were adopted throughout the AI-development process. Therefore, it is important to treat the process of AI development as circular, in that postdeployment evaluations could potentially promote adjustments in the first three stages. Like many of our recommendations for the other stages, adherence to these recommendations would require effort from various parties: lawmakers and policymakers, who could help enforce evaluations before and after deployment; owners of business corporations, who hold enormous power in shaping the industrial culture and distributing resources; and various developers, who are tasked with making assessments.

Mobilizing appropriate and required collective action to respond to structural injustice is never easy work. Rooted in structural-injustice analysis, our recommendations aim to provide a core theoretical framework upon which other domain experts can build to develop and realize more detailed mechanisms. With these collective efforts, we can better approach the goal of AI fairness.

7. Conclusion

The issue of AI bias poses urgent ethical challenges for modern societies and demands that efforts should be poured into pursuing AI fairness. However, to do so effectively, we must clarify what the notions of AI bias and AI fairness should entail. This paper argues that structural injustice provides a fruitful conceptual framework for analyzing the moral problem of AI bias, understanding the corresponding responsibility, and exploring more holistic responses to AI fairness. Viewed from this perspective, the problem of AI bias is a case of structural injustice resulting from interactions between AI and many other social factors. Furthermore, the goal of AI fairness should be to pursue a more just social structure, potentially with the development and use of AI systems when appropriate. Drawing on the SCM of responsibility, we further argue that all participating agents in the unjust social structure associated with AI bias bear a shared responsibility to join collective action with the goal of reforming the social structure. Accordingly, we provide a list of practical recommendations for agents in various social positions to contribute to this collective action.

Acknowledgements

Earlier versions of this paper were presented at the Feminist, Social Justice, and AI workshop, Department of Philosophy at the University of Georgia, Humanizing Machine Intelligence (HMI) project at Australian National University, Center for Bioethics at New York University, Department of Philosophy at University of Massachusetts Amherst, and the Institute of Philosophy of Mind and Cognition at National Yang Ming Chiao Tung University. We are very grateful to participants at the abovementioned venues as well as two anonymous referees of *Feminist Philosophy Quarterly* for constructive feedback.

References

Abràmoff, Michael D., Philip T. Lavin, Michele Birch, Nilay Shah, and James C. Folk. 2018. "Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices." *npj Digital Medicine* 1 (August): 39. https://doi.org/10.1038/s41746-018-0040-6.

Adamson, Adewole S., and Avery Smith. 2018. "Machine Learning and Health Care Disparities in Dermatology." *JAMA Dermatology* 154, no. 11 (November): 1247–48. https://doi.org/10.1001/jamadermatol.2018.2348.

Allen, K. D., C. G. Helmick, T. A. Schwartz, R. F. DeVellis, J. B. Renner, and J. M. Jordan. 2009. "Racial Differences in Self-Reported Pain and Function among Individuals with Radiographic Hip and Knee Osteoarthritis: The Johnston County Osteoarthritis Project." *Osteoarthritis and Cartilage* 17, no. 9 (September): 1132–36. https://doi.org/10.1016/j.joca.2009.03.003.

Anderson, Elizabeth. 1995. "Knowledge, Human Interests, and Objectivity in Feminist Epistemology." *Philosophical Topics* 23, no. 2 (Fall): 27–58.

Angwin, Julia, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2016. "Machine Bias." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aragon, Corwin, and Alison M. Jaggar. 2018. "Agency, Complicity, and the Responsibility to Resist Structural Injustice." *Journal of Social Philosophy* 49, no. 3 (Fall): 439–60. https://doi.org/10.1111/josp.12251.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104, no. 3 (June): 671–732. https://doi.org/10.15779/Z38BG31.

Chen, Irene Y., Shalmali Joshi, and Marzyeh Ghassemi. 2020. "Treating Health Disparities with Artificial Intelligence." *Nature Medicine* 26, no. 1 (January): 16–17. https://doi.org/10.1038/s41591-019-0649-2.

Chen, Irene Y., Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. "Ethical Machine Learning in Healthcare." *Annual*

Review of Biomedical Data Science 4:123–44. https://doi.org/10.1146/annurev-biodatasci-092820-114757.

Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. 2019. "How to Develop Machine Learning Models for Healthcare." *Nature Materials* 18, no. 5 (May): 410–14. https://doi.org/10.1038/s41563-019-0345-0.

Coeckelbergh, Mark. 2020. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26, no. 4 (August): 2051–68. https://doi.org/10.1007/s11948-019-00146-8.

Collins, Patricia Hill. 1990. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment.* New York: Routledge.

Dahl, Robert A. 1989. *Democracy and Its Critics*. New Haven, CT: Yale University Press.

Dastin, Jeffrey. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." Reuters, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Davis, Jenny L., Apryl Williams, and Michael W. Yang. 2021. "Algorithmic Reparation." *Big Data and Society* 8 (2). https://doi.org/10.1177/20539517211044808.

Eberly, Lauren, Dustin Richter, George Comerci, Justin Ocksrider, Deana Mercer, Gary Mlady, Daniel Wascher, and Robert Schenck. 2018. "Psychosocial and Demographic Factors Influencing Pain Scores of Patients with Knee Osteoarthritis." *PloS One* 13 (4): e0195075. https://doi.org/10.1371/journal.pone.0195075,

Escobar, Oliver, and Stephen Elstub. 2017. "Forms of Mini-Publics: An Introduction to Deliberative Innovations in Democratic Practice." *newDemocracy*, Research and Development Note 4, May 8, 2017. https://www.newdemocracy.com.au/2017/05/08/forms-of-mini-publics/.

Eveleth, Rose. 2014. "How Self-Tracking Apps Exclude Women." *Atlantic*, December 15, 2014. https://www.theatlantic.com/technology/archive/2014/12/how-self-tracking-apps-exclude-women/383673/.

Fallin-Bennett, Keisa. 2015. "Implicit Bias against Sexual Minorities in Medicine: Cycles of Professional Influence and the Role of the Hidden Curriculum." *Academic Medicine: Journal of the Association of American Medical Colleges* 90, no. 5 (May): 549–52. https://doi.org/10.1097/ACM.0000000000000662.

Farooq, Faheem, and John J. Strouse. 2018. "Disparities in Foundation and Federal Support and Development of New Therapeutics for Sickle Cell Disease and Cystic Fibrosis." *Blood* 132, Supplement 1 (November 29): 4687. https://doi.org/10.1182/blood-2018-99-115609.

Fazelpour, Sina, and Zachary C. Lipton. 2020. "Algorithmic Fairness from a Non-ideal Perspective." In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63. New York: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375828.

FDA (Food and Drug Administration). 2018. "FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems." FDA new release, April 11, 2018. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.

Fisk, N. M., and R. Atun. 2009. "Systematic Analysis of Research Underfunding in Maternal and Perinatal Health." *BJOG: An International Journal of Obstetrics and Gynaecology* 116, no. 3 (February): 347–56. https://doi.org/10.1111/j.1471-0528.2008.02027.x,

Fraser, Nancy, and Axel Honneth. 2003. *Redistribution or Recognition? A Political-Philosophical Exchange*. London: Verso.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. "On the (Im)Possibility of Fairness." arXiv preprint, arXiv:1609.07236 [cs.CY]. https://doi.org/10.48550/arXiv.1609.07236.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning." *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–38. New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287589.

Geiger, H. Jack. 2003. "Racial and Ethnic Disparities in Diagnosis and Treatment: A Review of the Evidence and a Consideration of Causes." In *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*, edited by Brian D. Smedley, Adrienne Y. Stith, and Alan R. Nelson, 417–54. Washington, DC: National Academies Press.

Good, Mary-Jo Delvecchio, Cara James, Byron J. Good, and Anne E. Becker. 2005. "The Culture of Medicine and Racial, Ethnic, and Class Disparities in Healthcare." In *The Blackwell Companion to Social Inequalities*, edited by Mary Romero and Eric Margolis, 396–423. Malden, MA: Blackwell.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, et al. 2016. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." *JAMA: The Journal of the American Medical Association* 316, no. 22 (December 13): 2402–10. https://doi.org/10.1001/jama.2016.17216.

Gunnemyr, Mattias. 2020. "Why the Social Connection Model Fails: Participation Is Neither Necessary nor Sufficient for Political Responsibility." *Hypatia* 35, no. 4 (Fall): 567–86. https://doi.org/10.1017/hyp.2020.40.

Hajian, Sara, and Josep Domingo-Ferrer. 2013. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining." *IEEE Transactions on Knowledge*

and Data Engineering 25, no. 7 (July): 1445–59. https://doi.org/10.1109/tkde.2012.72.

Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Noûs* 34, no. 1 (March): 31–55. https://doi.org/10.1111/0029-4624.00201.

Haslanger, Sally. 2016. "What Is a (Social) Structural Explanation?" *Philosophical Studies* 173, no. 1 (January): 113–30. https://doi.org/10.1007/s11098-014-0434-5.

Hedden, Brian. 2021. "On Statistical Criteria of Algorithmic Fairness." *Philosophy and Public Affairs* 49, no. 2 (Spring): 209–31. https://doi.org/10.1111/papa.12189.

Heidari, Hoda, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. "A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity." In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 181–90. New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287584.

Hellman, Deborah. 2020. "Measuring Algorithmic Fairness." *Virginia Law Review* 106, no. 4 (June): 811–66.

Hing, Esther, and Catharine W. Burt. 2009. "Are There Patient Disparities When Electronic Health Records Are Adopted?" *Journal of Health Care for the Poor and Underserved* 20, no. 2 (May): 473–88. https://doi.org/10.1353/hpu.0.0143.

Hoffmann, Anna Lauren. 2019. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication and Society* 22 (7): 900–915. https://doi.org/10.1080/1369118X.2019.1573912.

Hooker, Sara. 2021. "Moving beyond 'Algorithmic Bias Is a Data Problem.'" *Patterns* 2, no. 4 (April 9): 100241. https://doi.org/10.1016/j.patter.2021.100241.

hooks, bell. 1990. *Yearning: Race, Gender, and Cultural Politics*. Boston: South End Press.

Johnson, Gabbrielle M. 2021. "Algorithmic Bias: On the Implicit Biases of Social Technology." *Synthese* 198, no. 10 (October): 9941–61. https://doi.org/10.1007/s11229-020-02696-y.

Joyce, Kelly, Laurel Smith-Doerr, Sharla Alegria, Susan Bell, Taylor Cruz, Steve G. Hoffman, Safiya Umoja Noble, and Benjamin Shestakofsky. 2021. "Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change." *Socius* 7. https://doi.org/10.1177/2378023121999581.

Kamiran, Faisal, Toon Calders, and Mykola Pechenizkiy. 2013. "Techniques for Discrimination-Free Predictive Models." In *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, edited by Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky, 223–39. New York: Springer. https://doi.org/10.1007/978-3-642-30487-3_12.

Kaushal, Amit, Russ Altman, and Curt Langlotz. 2020. "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms." *JAMA: The Journal of the American Medical Association* 324, no. 12 (September 22/29): 1212–13. https://doi.org/10.1001/jama.2020.12067.

Kelly, Christopher J., Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. "Key Challenges for Delivering Clinical Impact with Artificial Intelligence." *BMC Medicine* 17 (October 29): 195. https://doi.org/10.1186/s12916-019-1426-2.

Köhler, Sebastian, Neil Roughley, and Hanno Sauer. 2017. "Technologically Blurred Accountability? Technology, Responsibility Gaps and the Robustness of Our Everyday Conceptual Scheme." In *Moral Agency and the Politics of Responsibility*, edited by Cornelia Ulbert, Peter Finkenbusch, Elena Sondermann, and Tobias Debiel, 51–68. New York: Routledge.

Le Bui, Matthew, and Safiya Umoja Noble. 2020. "We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness." *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 162–79. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.9.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521, no. 7553 (May 28): 436–44. https://doi.org/10.1038/nature14539.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes." *Philosophy & Technology* 31, no. 4 (December): 611–27. https://doi.org/10.1007/s13347-017-0279-x.

Liao, Shen-Yi, and Vanessa Carbonell. 2022. "Materialized Oppression in Medical Tools and Technologies Materialized Oppression in Medical Tools and Technologies." *American Journal of Bioethics*. Published online ahead of print, March 9, 2022. https://doi.org/10.1080/15265161.2022.2044543.

Lin, Ting-An. Forthcoming. "Sexual Violence and Two Types of Moral Wrongs." *Hypatia.*

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Lu, Catherine. 2017. *Justice and Reconciliation in World Politics*. Cambridge: Cambridge University Press.

Manuel, Jennifer I. 2018. "Racial/Ethnic and Gender Disparities in Health Care Use and Access." *Health Services Research* 53, no. 3 (June): 1407–29. https://doi.org/10.1111/1475-6773.12705.

Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6, no. 3 (September): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

McKeown, Maeve. 2016. "Global Structural Exploitation: Towards an Intersectional Definition." *Global Justice: Theory Practice Rhetoric* 9 (2): 155–77. https://doi.org/10.21248/gjn.9.2.116.

———. 2018. "Iris Marion Young's 'Social Connection Model' of Responsibility: Clarifying the Meaning of Connection." *Journal of Social Philosophy* 49, no. 3 (Fall): 484–502. https://doi.org/10.1111/josp.12253.

Miconi, Thomas. 2017. "The Impossibility of 'Fairness': A Generalized Impossibility Result for Decisions." arXiv preprint, arXiv:1707.01195 [stat.AP]. https://doi.org/10.48550/arXiv.1707.01195.

Nelson, Alan. 2002. "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care." *Journal of the National Medical Association* 94, no. 8 (August): 666–68.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Nussbaum, Martha C. 2009. "Iris Young's Last Thoughts on Responsibility for Global Justice." In *Dancing with Iris: The Philosophy of Iris Marion Young*, edited by Ann Ferguson and Mechthild Nagle, 133–45. Oxford: Oxford University Press.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366, no. 6464 (October 25): 447–53. https://doi.org/10.1126/science.aax2342.

Pierson, Emma, David M. Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. 2021. "An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations." *Nature Medicine* 27, no. 1 (January): 136–40. https://doi.org/10.1038/s41591-020-01192-7.

Poleshuck, Ellen L., and Carmen R. Green. 2008. "Socioeconomic Disadvantage and Pain." *Pain* 136, no. 3 (June): 235–38. https://doi.org/10.1016/j.pain.2008.04.003.

Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. "Ensuring Fairness in Machine Learning to Advance Health Equity." *Annals of Internal Medicine* 169, no. 12 (December 18): 866–72. https://doi.org/10.7326/M18-1990.

Rodriguez, Rudolph A., Saunak Sen, Kala Mehta, Sandra Moody-Ayers, Peter Bacchetti, and Ann M. O'Hare. 2007. "Geography Matters: Relationships among Urban Residential Segregation, Dialysis Facilities, and Patient Outcomes." *Annals of Internal Medicine* 146, no. 7 (April 3): 493–501. https://doi.org/10.7326/0003-4819-146-7-200704030-00005.

Rudy-Hiller, Fernando. 2018. "The Epistemic Condition for Moral Responsibility." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018

edition. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/.

Schwenkenbecher, Anne. 2021. "Structural Injustice and Massively Shared Obligations." *Journal of Applied Philosophy* 38, no. 1 (February): 23–39. https://doi.org/10.1111/japp.12431.

Shankar, Shreya, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World." arXiv preprint, arXiv:1711.08536 [stat.ML]. https://doi.org/10.48550/arXiv.1711.08536.

Smith, Aaron. 2018. "Public Attitudes toward Computer Algorithms." Pew Research Center, November 16, 2018. https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/.

Tigard, Daniel W. 2021. "There Is No Techno-Responsibility Gap." *Philosophy & Technology* 34, no. 3 (September): 589–607. https://doi.org/10.1007/s13347-020-00414-7.

van de Poel, Ibo, Jessica Nihlén Fahlquist, Neelke Doorn, Sjoerd Zwart, and Lambèr Royakkers. 2012. "The Problem of Many Hands: Climate Change as an Example." *Science and Engineering Ethics* 18, no. 1 (March): 49–67. https://doi.org/10.1007/s11948-011-9276-0.

van der Veer, Sabine N., Lisa Riste, Sudeh Cheraghi-Sohi, Denham L. Phipps, Mary P. Tully, Kyle Bozentko, Sarah Atwood, et al. 2021. "Trading off Accuracy and Explainability in AI Decision-Making: Findings from 2 Citizens' Juries." *Journal of the American Medical Informatics Association: JAMIA* 28, no. 10 (October): 2128–38. https://doi.org/10.1093/jamia/ocab127.

Vidyasagar, D. 2006. "Global Notes: The 10/90 Gap Disparities in Global Health Research." *Journal of Perinatology* 26, no. 1 (January): 55–56. https://doi.org/10.1038/sj.jp.7211402.

Vyas, Darshali A., Leo G. Eisenstein, and David S. Jones. 2020. "Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms." *New England Journal of Medicine* 383, no. 9 (August 27): 874–82. https://doi.org/10.1056/nejmms2004740.

Wawira Gichoya, Judy, Liam G. McCoy, Leo Anthony Celi, Marzyeh Ghassemi. 2021. "Equity in Essence: A Call for Operationalising Fairness in Machine Learning for Healthcare." *BMJ Health & Care Informatics* 28 (1): e100289. https://doi.org/10.1136/bmjhci-2020-100289.

Wong, Pak-Hang, and Judith Simon. 2020. "Thinking About 'Ethics' in the Ethics of AI." *Idees*, no. 48 ("Artificial Intelligence"). https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/.

Wylie, Alison, and Lynn Hankinson Nelson. 2007. "Coming to Terms with the Values of Science: Insights From Feminist Science Studies Scholarship." In *Value-Free*

Science? Ideals and Illusions, edited by Harold Kincaid, John Dupré, and Alison Wylie, 58–86. New York: Oxford University Press.

Young, Iris Marion. 2006. "Responsibility and Global Justice: A Social Connection Model." *Social Philosophy & Policy* 23, no. 1 (January). Cambridge University Press: 102–30. https://doi.org/10.1017/S0265052506060043.

———. 2011. *Responsibility for Justice*. Oxford: Oxford University Press.

Zhao, Qingyu, Ehsan Adeli, and Kilian M. Pohl. 2020. "Training Confounder-Free Deep Learning Models for Medical Applications." *Nature Communications* 11:6010. https://doi.org/10.1038/s41467-020-19784-9.

Zheng, Robin. 2018. "What Is My Role in Changing the System? A New Model of Responsibility for Structural Injustice." *Ethical Theory and Moral Practice* 21, no. 4 (August): 869–85. https://doi.org/10.1007/s10677-018-9892-8.

Zimmermann, Annette, Elena Di Rosa, and Hochan Kim. 2020. "Technology Can't Fix Algorithmic Injustice." *Boston Review*, January 9, 2020. https://bostonreview.net/articles/annette-zimmermann-algorithmic-political/.

TING-AN LIN is an interdisciplinary ethics postdoctoral fellow in the McCoy Family Center for Ethics in Society and the Institute for Human-Centered Artificial Intelligence at Stanford University. She earned her PhD in philosophy from Rutgers University, where she also received a graduate certificate in women's and gender studies. She specializes in ethics, feminist philosophy, and social and political philosophy. Her recent research focuses on ethical issues that arise from the interactions between social structures and individuals, and she is developing a moral framework for addressing them.

PO-HSUAN CAMERON CHEN is a staff software engineer and a tech lead manager of machine learning at Google Research and Google Health. Cameron's primary research interests lie at the intersection of machine learning and health care. His research has been published in leading scientific, clinical, and machine learning venues, including *Nature*, *JAMA*, and *NeurIPS*. He received his PhD in electrical engineering and neuroscience from Princeton University and his BS in electrical engineering from National Taiwan University. This work was done in personal time; views are those of the authors and do not necessarily reflect the official policy or position of Google LLC.