The Reach of Fairness[1]

Joshua Cohen

FACULTY SENIOR DIRECTOR, Apple University; DISTINGUISHED SENIOR FELLOW IN LAW, PHILOSOPHY, AND POLITICAL SCIENCE, UC Berkeley; CO-EDITOR, Boston Review, jcohen57@berkeley.edu

Lydia T. Liu

ASSISTANT PROFESSOR OF COMPUTER SCIENCE, Princeton University, ltliu@princeton.edu

In conversation with Barocas, Hardt, and Narayanan's *Fairness and Machine Learning* (FaML), we seek to broaden the scope of normative argument about machine learning and algorithmic decision making. Beginning from an understanding of fair cooperation among free and equal persons as a fundamental political value, we argue that concerns about fairness and machine learning need to be expanded in three ways. First, unfairness and discrimination are not only a matter of systematic group subordination. We consider other forms of unfairness that are not about disadvantaged groups but about removing barriers to opportunity, and suggest practical implications for algorithmic decisions. Secondly, while we broadly agree with FaML's approach to fair organizational decisions, we underscore the limits of a focus on fair organizational decisions in advancing equality of opportunity in society. Finally, drawing on Rawls, we present aspects of a fair society that are not simply matters of equal opportunity, and consider some broader, under-explored ramifications of algorithms and AI on societal fairness. Specifically, we suggest the implications that AI deployment at scale has for fair distribution of income and wealth, political liberties, and public deliberation.

CCS CONCEPTS • Machine Learning • Artificial Intelligence

Additional Keywords and Phrases: Algorithmic Fairness, Political Philosophy, Societal Impacts, Equality of Opportunity and Algorithmic Decision Making, Rawlsian Justice, Broader Impacts of Large Language Models

ACM Reference Format:

1 INTRODUCTION

Machine-executed algorithms[2] (hereafter, "algorithms") are doing lots of work in our lives.[3] Algorithms play a role in search, games, ads, news, romance, books, movies, restaurants, and music; they provide input into decisions about policing, bail, credit, hiring, and admissions; and they figure in high-frequency trading, organ donations, and histopathology.

Now add generative AI, rinse, exponentiate.

Much good can come from these innovations. But they also raise serious concerns, as when face recognition systems "mis-gender women and darker skinned individuals"[4]; video recommendations land in fever swamps; deep fakes misrepresent a candidate's likeness and words; reputations are damaged by "large libel models"[5]; micro-targeted political messages turn public discussion into personalized marketing opportunities; training data are appropriated without compensation; or creative destruction turns into the destruction of creatives.

Are these troubles the broken eggs that herald great omelets? Old injustices wrapped in new technologies? The portentous opening scenes of humanity's last act?

These questions have sparked wide-ranging debate, including debates about fairness in machine learning and AI. Much of this work is alert to the benefits that machine learning can bring to organizational decisions—aiding efficiency, constraining arbitrariness, identifying patterns unnoticed by human decision-makers. But it is especially concerned about risks of unfairness, with a particular focus on unfair organizational decisions about individuals—decisions about jobs, university admissions, credit, insurance, bail, and parole.

In this article, we offer some critical reflections on standard approaches to fairness and machine learning. We focus on *Fairness and Machine Learning* (FaML), a recent book by Solon Barocas, Moritz Hardt, and Arvind Narayanan (hereafter BHN). FaML provides a strikingly thorough, synthetic, technically sophisticated, and normatively rich treatment of these issues. Though we focus on FaML, we have an eye to addressing a broader set of themes.

FaML focuses principally—as the title indicates—on risks of unfairness (though it is also attentive to concerns about the *legitimacy* of organizational uses of machine learning in making consequential decisions).[6] As we mentioned, much of the existing work on fairness and machine learning—and this is true of FaML—focuses more particularly on risks of unfairness in

organizational (often private organizational) decision-making.[7] Thus BHN identify unfairness and discrimination,[8] and associate discrimination/unfairness with the reinforcement of "systematic" group disadvantage (83-84)—that is, disadvantage or subordination for groups across a broad range of social goods and opportunities (84).[9] They then train a critical eye on the use of machine learning by organizations in ways that differentially treat people "according to characteristics like race, gender, and disability." (85) Such uses are understood as unfair because they reproduce systematic group disadvantage and undermine equality of opportunity.

To contextualize BHN's approach and our subsequent reflections, it is instructive to consider prior philosophical and analytical work on fairness in machine learning.[10] Scholars such as Hedden (2021), Green (2022), Fazelpour and Lipton (2020), and Fazelpour, Lipton, and Danks (2022) have criticized the conceptual shortcomings of purely formal approaches to fairness in machine learning (also known as *algorithmic fairness*), emphasizing the need for substantive evaluations of algorithmic impacts on societal fairness.[11] In addition, technical analyses by Liu et al. (2018) and Corbett-Davies et al. (2023) showed that formal fairness criteria can sometimes fail to achieve intended outcomes when considering the desired allocative impact and downstream effects of algorithmic decisions.[12]

BHN avoid these shortcomings by going beyond the mathematical formalization of fairness requirements and the exploration of their inconsistencies in prior work.[13] In particular, BHN emphasize that substantive fairness resists formalization, in part because fair decisions require context-specific, contestable judgments, including judgments about which groups are owed special concern and why (83). As a corollary, they resist the temptation to interpret fair decision-making very abstractly, as a matter of computationally eliminating any disparities—say, disparities in error rates across all recognizable subsets of the population—from machine learning models.[14] And they recognize social inequalities that lie beyond the reach of organizational decisions (229-38). But they focus principally on organizational decision-making, aiming to strike a delicate—sometimes perplexing—balance between fair uses of machine learning in organizational decisions and broader concerns about its implications for a fair society.

The fact that organizations sometimes make decisions that reproduce group disadvantage is of course not news, nor simply a product of machine learning. And like much of the literature, FaML does not claim that machine learning worsens unfairness or group disadvantaging relative to a pre-machine-learning baseline.[15] As the "opportunities" in their subtitle indicates, BHN convey a sense of possibility that mechanical procedures trained on large quantities of data might reduce the role of human biases and cognitive limitations, thus enhancing the fairness of organizational decisions. They are "cautiously optimistic about fairness and machine learning" (23). But as the "limitations" in their subtitle indicates, the principal aim of FaML is to underscore the large challenges in delivering on these important ambitions.

We share this mix of hopeful sensibility and cautionary concern. In the interest of advancing the discussion beyond mere agreement, we will focus the remainder of the paper elsewhere. Premising that fairness is a fundamental value in political morality, we will consider three ways in which current discussions of fairness and machine learning—including those in FaML—are truncated.

First, a common idea is that the "normative force" of discrimination—what makes it wrong—comes from the wrongness of group disadvantage. BHN, for example, say that the "normative force" of discrimination flows from the wrongness of group disadvantage (84). Normatively objectionable discrimination—unfair treatment[16]—is thus "treatment that systematically imposes a disadvantage on one social group relative to others" (83-84). This is too narrow. *Unfair organizational decisions are not exclusively about systematic group subordination.*

Second, the discussion of fairness and machine learning often associates fairness with the value of equal opportunity. But *achieving fair equal opportunity in a society lies well beyond the reach of organizational decisions.* It is implausible to think that organizational decisions can rectify the background social inequalities in which organizations operate.[17] BHN fully agree, so fully that they wonder "why should we continue to study the notion of discrimination in decision-making" (xii, 20).[18] Our intention in raising this second concern, then, is not to point to an oversight, but to underscore some fundamental issues that lie outside the focus of FaML but are important parts of a broader discussion of normative concerns about machine learning.

Third, while equality of opportunity is a requirement of fairness, fairness *has far broader reach than equality of opportunity*. Consider John Rawls' theory of justice. Drawing on an idea of fair social cooperation[19]—specifically, cooperation among free and equal persons[20]—he argues for a conception of justice that includes equality of opportunity but also requires protections for basic liberties, including the political liberties associated with democracy, and a distribution of resources that

maximizes advantage for the least advantaged. One need not agree with the specifics of Rawls' principles of justice to feel the force of this expansive understanding—beyond equal opportunity—of fair terms of social cooperation.

The unifying thread tying these three issues together is that the normative implications of machine learning (and other forms of algorithmic system) for fairness need to be explored in a broader political register.[21] To be sure, FaML's focus on organizational decisions and group subordination is understandable. BHN concentrate on the United States (xiii), where self-regulation by organizations has been the main way of addressing concerns about machine learning. Consider, in contrast, the European Union's GDPR, Digital Services Act, Digital Markets Act, and AI Act, all of which—whatever the precise merits of the legislation—are more broadly concerned with issues about fundamental rights and democracy. As the EU examples indicate, normative concerns about machine learning and automated recommendations and decisions have very broad reach.

If fairness is our fundamental value in exploring the normative implications of machine learning, then we should be attentive to—here are our three themes—unfair treatment that is not about group subordination; sources of unfairness that lie beyond the reach of organizational decisions; and norms of fair social cooperation that lie beyond ensuring equal opportunity. In short, our main aim is to present a broader framework for exploring normative concerns about fairness and machine learning—a set of *principles*, rooted in the core value of fairness, that might guide public judgment about consequential applications.

2 DISCRIMINATION AND SYSTEMATIC SUBORDINATION

BHN say that a group is *subordinated* just in case it is subject to multi-dimensional or "systematic" disadvantages—say, lifetime disadvantages in health, education, income, wealth, and neighborhood stressors. Moreover, this "systematicity in the differences in treatment and outcomes is what gives discrimination its normative force as a concept" (84). An organization wrongfully discriminates, then, if *and only if* the organization makes decisions that reflect and help to reproduce systematic group disadvantage. BHN thus reject the "anti-classification" view of discrimination according to which organizations engage in wrongful discrimination simply by using "suspect" classifications in their decisions—say, classifying people according race or gender or national origin in making hiring or university admissions decisions, or in awarding subcontracts or making loans. Using those classifications is not wrongful discrimination, BHN think, when they are used to disrupt patterns of systematic group disadvantage.

We agree in rejecting the anti-classification view. Sometimes fairness is aided by awareness.[22] And we agree that disrupting systematic group disadvantage is of great social importance. But organizational decisions can be wrongfully discriminatory without sustaining group disadvantage. BHN's understanding of organizational fairness and anti-discrimination is too narrow. There are compelling reasons for concern about unfair decisions even when groups are not at issue, or when there is a group but it is not systematically disadvantaged. If we are right, then designers of ML algorithms should not confine their concerns about the discriminatory effects of those algorithms to issues of group subordination.

2.1. Consider the US employment context. With employment typically at will, employers have very sweeping discretion in decisions about hiring, firing, promotion, and job requirements. Anti-discrimination law bounds that discretion, thus providing some protection for opportunity. As a legal matter, one anti-discrimination requirement is that employers accommodate religious beliefs and practices. Standard claims for religious accommodation involve attire and appearance, on-site worship, and scheduling (respecting sabbatarian practices). Title 7 of the Civil Rights Act of 1964 requires employers to make reasonable accommodations for religious practices, unless the accommodation imposes an "undue hardship" on the employer.[23] Understandings of "undue hardship" have differed over time and across institutions—including the Equal Employment Opportunity Commission (EEOC), courts, and state legislatures. In a 1977 case, *Trans World Airlines v. Hardison*, the Supreme Court suggested (over the dissents of Justices Marshall and Brennan) that an accommodation that imposes more than a "de minimis" cost on an employer is an undue hardship. But the Hardison standard was never really consistently followed by lower courts or by the EEOC, nor taken as a beacon by state legislatures. In the 2023 case of *Groff v. DeJoy*, the Supreme Court unanimously set it aside.[24]

Our aim here is not to discuss the legal issues about religious accommodations. Instead, we use them to illustrate the normative significance of claims of religious discrimination. Failures to provide religious accommodations are not matters of systemic group subordination. Dissenting in Hardison, Justice Thurgood Marshall forcefully stated the essential normative concern: "a society that truly values religious pluralism cannot compel adherents of minority religions to make the cruel choice of surrendering their religion or their job."[25]

Why is the forced choice cruel? Not because the adherents in question—whether Seventh Day Adventists, ultra-orthodox Jews, or]evangelical Christians[26]—are members of systematically disadvantaged groups. To be sure, some people who make claims for religious accommodations may present themselves as members of culturally, socially, politically beleaguered minorities. But we need not embrace that arguably extravagant self-understanding to find the claim for an accommodation compelling, and to think that the failure to provide it is a normatively important kind of discrimination/unfairness.

A person who complains about being unreasonably denied an accommodation thinks they are being treated wrongly, and may present the wrongness as *unfairness—a non-comparative* unfairness.[27] The complaint is not that they are being treated less well than comparably-situated others, but that their interest in fulfilling their religious obligations, as they understand them, is not being given due weight, and that they are therefore not being given the treatment that they are owed. The non-comparative case for an accommodation grows from the supreme and practice-guiding importance of sincerely held religious convictions in the life of the believer. For example, the sabbatarian believes, as part of a more or less comprehensive structure of convictions, that they are required to rest on the sabbath—required not by a voluntarily undertaken obligation, but by God's law. Or the Muslim makes a case for an accommodation for additional time for on-site prayer during Ramadan.

To be sure, you can try to interpret a failure to accommodate as a group disadvantage. But claimants do not call for the accommodation as members of a group, and certainly not as a remedy for a group disadvantage,[28] but as bearers of the conviction. Moreover, there is no reason to suppose that the people who share the conviction are *systemically* disadvantaged, or even unable to find other employers that might provide accommodations. The objectionableness of the treatment is simpler and forcefully identified by Marshall: they are being required, and not for a sufficiently good reason, to choose between their job and their religion. It is the deep importance of the convictions, not a pervasive group disadvantage, that demands the accommodation.

2.2. The religious accommodations case suffices to make the general point about claims of discrimination/unfairness that are not matters of group subordination. But it is by no means the only example. Lots of states in the US have anti-discrimination laws that protect at least some speech or political engagement by private employees, not because of membership in a protected class, but because of the fundamental importance of the activity.[29] And in his illuminating book *Bottlenecks*, Joseph Fishkin describes a range of employment discrimination laws adopted by US states that seek to reduce hurdles to opportunity—what he calls "bottlenecks"—that are not plausibly construed as targeted on systemic group subordination.[30] His principal examples are laws preventing employers from asking on job applications about current employment status, credit history, or criminal convictions. The idea in these laws is to help people get past a first hurdle in applying for a job.[31]

To be sure, given racial disparities in unemployment, incarceration, income, and wealth, these laws can be construed as indirectly targeted on race discrimination. But on their face, they are not indirectly targeted at racial disparities, but at hurdles to opportunity that are shared by members of protected classes and people who are not in those classes. People who are not members of protected classes benefit from these laws, and not because they are the unintended beneficiaries of laws that are imperfectly targeted corrections for group subordination. Instead, they are the intended beneficiaries of laws targeted on unfair hurdles to opportunity.

2.3. So even within the context of organizational decisions, the focus on systemic group subordination is too narrow as it excludes other compelling cases such as religious accommodation and removing barriers to opportunity that are not primarily about systematic group disparities. How might normative discussions of machine-learning algorithms in the employment context look different under an expanded understanding of anti-discrimination?

Consider the use of algorithms to provide an initial screen of job applicants. Suppose we find that the algorithm screens out all job applicants who are currently unemployed (or have a low credit score, or prior criminal conviction), but we also find no disparate impact of the algorithm across dimensions of historical group subordination such as race, gender, and disability status. Or suppose the screening algorithm provides low rankings to candidates with checkered employment records because of conflicts with employers growing from a failure of previous employers to provide religious accommodations. While current unemployment may signal concerns about qualifications, there can well be qualified job candidates who are currently unemployed, and subject to objectionable discrimination and exclusion from opportunities.

Our discussion suggests that even if the adjustment has no effect on systematically subordinated groups, we still have a very strong argument on grounds of fairness for adjusting the algorithm to allow some of the systematically denied candidates to pass through the initial screening.

3 ORGANIZATIONAL FAIRNESS AND EQUALITY OF OPPORTUNITY?

BHN, as we have said, are principally focused on the fairness of organizational decisions, and closely associate fairness with equal opportunity. In particular, they link their conception of fair organizational decision-making with what they call the *middle view of equality of opportunity*, which they distinguish from *narrow* and *broad* views.

3.1. The *narrow* view focuses on organizational decisions, and takes people as they are. In the employment context, for example, it says that organizations provide equal opportunity when they hire the most qualified applicants, however they acquired those qualifications. The *broad* view differs from the narrow view in two ways. It is not about organizational decisions, but about the social structure (including laws and public policies). One version of the broad view might say that a society provides *equality of opportunity* when people with similar abilities and aspirations having equal life chances, regardless of morally irrelevant differences in class, gender, race, and other qualities that differentiate free and equal persons.[32] And because the broad view is focused on the social structure, not on organizational decisions, it does not take people as they are. Instead, it is concerned to ensure that people have fair chances to acquire the qualifications that enable them to pursue their aspirations.

The *middle view* is a hybrid. Like the narrow view, it is "narrowly concerned" with organizational decisions (92). But like the broad view, it does not take people as they are. It says that organizations make fair decisions when they evaluate people in part by considering how the organization's own decisions might help to sustain or limit the reproduction of systematic group disadvantages. And BHN suggest a counterfactual test for implementing this view: "the middle view . . . suggests that ensuring equality of opportunity requires assessing people as they *would have been* had they been offered in the past opportunities comparable to other people of equal potential seeking the current opportunity" (92; also 93).

In effect, the narrow view treats the relationships between employer and employee, hospital and patient, lender and borrower, university and applicants, judge and accused as *isolated transactions*, to be assessed by standards of transactional justice. The middle view, in contrast, sees those transactions as part of a broader pattern of social relationships, including patterns of group subordination, of which they are parts. Organizational decisions—whom to hire, admit as a student, lend to, or hold without bail—are sites of social reproduction, among the actions through which structures of group subordination are sustained. Thus the middle view, like the narrow view, is "concerned with the fairness of [organizational] decision-making," but it is, like the broad view, "sensitive to the dynamics by which disadvantage might be perpetuated in society more broadly" (91-92).

3.2. In the discussion that follows, we ask a specific question. We will assume—as BHN do (91)—a Rawlsian version of the broad view of equality of opportunity. Thus understood, the broad view is not about rectifying historical injustices. It is a forward-looking requirement about creating a fairer society, in which the availability of opportunities are freed from the accidents of birth, in which genesis is not destiny. We want to explore how much we advance that goal by ensuring the fairness of organizational decisions, guided by the middle view of equal opportunity. We will present two examples—stylized versions of real cases—that underscore the limited gains that come from the middle view. Our aim is not to criticize the middle view, understood as a guide for organizational decisions; we think that some version of the middle view is the right approach for many organizations. Instead we are underscoring the limits of the organizational focus as a way to advance broad equality of opportunity, and as a consequence, doing so by ensuring that organizations use allocatively fair algorithms.

The first example draws on David Robinson's compelling discussion in his *Voices in the Code* of the algorithm for allocating kidneys for transplant.[33] There are roughly 100,000 people on the allocation list, and roughly 10,000 transplants each year. The allocation algorithm was initially developed in the 1980s as an alternative to subjective, discriminatory, seat-of-the-pants judgments by doctors or hospital boards about the allocation of scarce kidneys. And it was subsequently revised through a long and, on Robinson's telling, genuinely thoughtful process of public deliberation aimed at achieving a fair balance of the different claims of people in need of new kidneys who are on the list of potential recipients.

As a result of these efforts, we have arguably achieved fairness in decisions about how to select people who are on the transplant list. But fairness to people on the list is not fairness in allocating kidneys. For that, there remains the question of who ends up on the list. And here there are, to borrow Fishkin's helpful term, some serious bottlenecks. While Medicare covers the costs of dialysis for its full duration, it only covers the costs of three years of the medications that recipients of kidney transplants need. As a result, some people stay on dialysis rather than trying to get a transplant because they are unable to cover the costs of medication. Others do not get the information or have the initial testing needed to get on the list. Those people are poorer and darker skinned than people who land on the list.

So how should we assess the fairness of the kidney allocation algorithm? Should we ask whether it fosters a fair allocation of kidneys, fair access to health care, fair access to health, or fair treatment for people on the list? The discussions of a fair algorithm for allocating kidneys were focused on the last of these. But *algorithmic* fairness for people on the list, as important as that is, will not correct for the unfairness in who gets on the list. So it would be wrong to conclude from the achievement of *fairness in organizational decision making* that we have achieved fairness in the allocation of kidney transplants. If our purpose is to to ensure fair access to transplants (or health care, or health), even an optimally fair algorithm—however it is defined—does not get us there.

It comes as no surprise that fairness in organizational decision making is a fundamentally limited approach for advancing equality of opportunity. As an organizational decision making system, the kidney allocation system is constrained by its own purpose and controls. For example, it does not control insurance coverage, nor can it make the aftercare for kidney transplants more affordable. In other words, it is unable to rectify the background unfairness in society—racial and socioeconomic disparities in health and access to healthcare—that inevitably shapes its allocations.

3.3. Our first example highlights the limits of fairness in organizational decisions in achieving fair equality of opportunity, but it may be an especially unusual case because of the literal limits on who gets on the list. So consider a second example, concerning the allocation of a scarce resource in hospitals.

Early in the Covid-19 pandemic, hospitals faced serious concerns about the availability of scarce ventilators. In response, some critical care units developed algorithms for ventilator allocation. They used these algorithms to estimate chances of how long patients would survive with ventilator access, and ranked candidates for ventilators based on these predictions.[34] The algorithms did not consider race, gender, income, religion, nationality, or any other objectionable basis for making life and death decisions. They focused instead on a concern that is relevant to the allocation of scare life-saving medical technology, viz. life expectancy, the years of life they would save. The scarce equipment should be used, according to the judgment of the critical care specialists, where it will do the most good, as measured by years of life saved.

But as the vast literature on social determinants of health tells us, there are racial and class disparities in morbidity and mortality. Suppose we ask now whether the years-of-life-saved algorithm for allocating ventilators is fair. Once we know the background disparities, it is very hard to answer that question without considering the social determinants that cause current differences in qualification.

Suppose the hospital adjusts the years-of-life-saved algorithm for racial and class disparities in life expectancy, effectively making it more likely for members of a disadvantaged group to receive a ventilator than the original algorithm. For example, the University of Pittsburgh critical care department decided to assign equal priority to all patients who would live five years or more if they had ventilator access.[35] Here we have the middle view in action; part of the reason for the policy was to correct for differences in life expectancy resulting from sociodemographic determinants of health, while also trying to make effective use of scarce medical resources.

Why make such an adjustment? BHN suggest two reasons. First, "decision makers have an obligation to avoid perpetuating injustice" (92). The idea is to reduce complicity with social inequities by reducing their impact on current decisions. A second reason is that middle-view-inspired decisions have a particularly significant impact on historically disadvantaged groups, "bring[ing] about greater change than any one of the more continuous interventions required of the broad view." (93) Here, the idea is not about avoiding complicity in past injustices but about creating a future in which access to health is not dependent on race or class.

If we take the broad view of equal opportunity as our guide, we will focus on the second rationale: about creating a more just future. But even if the adjusted algorithm allocates more ventilators to a historically disadvantaged group, it does little to "uplift" the historically disadvantaged group's access to health and hence advance equality of opportunity. This intervention at the level of organizational decision making does not ameliorate the many social sources of racial and class disparities in health, such as environmental and social stressors. This discrepancy is the basis for Michael Marmot's provocation in *The Health Gap*, "Why treat people and send them back to the conditions that make them sick?"[36] As BHN say: "Fairly rendered decisions under unfair circumstances may do little to improve people's lives" (20). To be sure, doing *something* is a good thing, and treating people fairly in the hospital is better than treating them unfairly, even if they land back in the conditions that made them sick. But as with the kidney allocation algorithm, if the aim is fair access to health, we are working with very limited tools.

3.4. One final observation about the middle view. We mentioned earlier that BHN propose to operationalize the middle view through an act of moral (re-)imagination. How, for example, should the hospital have operationalized the middle view in the

allocation of ventilators? In BHN's view, the hospital should have performed a *counterfactual exercise* by "assessing people as they *would have been* had they been afforded the past opportunities comparable to other people of equal potential seeking the current opportunity" (92).

Making decisions based on a counterfactual calculation of a person's life expectancy had they gotten the same opportunities at health as another person is implausible for two reasons. First, there are conceptual puzzles associated with the counterfactual exercise: how is the hospital to judge what a patient's life expectancy would have been had the society been fair?[37]

Second, and perhaps more to the point, what is a hospital to do with the revelation, guided by insights into social determinants of health, that someone—who will not live past the next month, even with ventilator access—*would have had* many more years to live, had they resided, for example, in a less toxic and harsh neighborhood environment[38] than another patient with longer life expectancy. Is the hospital to allocate scarce life-saving resources based on such counterfactuals, while ignoring the reality of actual years of life saved? The hospital may not wish to be complicit in the unjust circumstances that have produced the health disparities. But the hospital also cannot correct these circumstances. And it seems irresponsible to assign a ventilator to someone with a month left to live rather than assigning to a person with five years.

But our principal concern about the middle view is not with the implausible operationalization through counterfactual assessments. Instead, our two examples are meant to express skepticism about the power of applying the middle view in making organizational decisions as a way to achieve broad equality of opportunity. This is not at all an objection to the advantages of the middle over the narrow view. Organizations ought to look beyond "qualification" narrowly conceived, say, as measured by conventional tests (or life expectancy at hospital admission). But if our aim is to achieve a fair society in which genesis does not determine destiny, organizational decisions guided by the middle view—as important as they are—do too little, too late.

In registering this limitation, we deliberately elide important questions about the proper *agents of justice*. Those questions are about who is responsible for carrying out the work of justice.[39] Our concern here is not on who *should* bear this responsibility, but rather on what capacity organizations have to contribute to justice—not so much whether they *should*, but on how much they can do. Our claim that organizations, especially private organizations, lack this broad capacity invites the following question: what constructive role could machine learning play in public health decision making aimed at better understanding social determinants of health and identifying and rectifying background health disparities in service of achieving fair equality of opportunity? That is an excellent question, one we hope to pursue elsewhere, and one that we urge more researchers—including machine learning researchers—to investigate beyond the scope of improving organizational decision making.

4 JUSTICE BEYOND EQUALITY OF OPPORTUNITY

Though "justice beyond fair decision making" is not the "core concern" in FaML (20), BHN emphasize the importance of the topic, and return to it throughout the book. For example, they ask whether the adoption of machine learning by decision makers "help[s] us make progress toward enabling equality of opportunity, or other normative ideals, over the course of people's lives" (229). But their focus on organizational decision making limits the exploration of these broader concerns—both concerns about equality of opportunity and about other normative ideals. Here we want to identify some issues that belong to a discussion of fairness and machine learning, but fall outside their main focus. For reasons of specificity and familiarity, and because we find it plausible, we will draw on John Rawls's theory of justice. We will begin with a brief sketch of his view, emphasizing the essential role of an idea of fairness, and then use it to gesture at some of those issues.

4.1. Many people, perhaps most readers of this journal, are drawn to the idea that fairness is not only an important virtue of persons and organizational decisions, but of our societies, which ought provide fair terms of social cooperation. Rawls's account of justice is animated by that idea and aims to say what fair terms of social cooperation among *free and equal persons* are. His proposal is captured in two principles of justice:

a. "Each person has an equal claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.
b. "Social and economic inequalities are to satisfy two conditions: first, they are to be attached to positions and offices open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society."[40]

These two principles include the broad conception of equal opportunity that BHN refer to, but also require that we have equal basic liberties and that inequalities in income and wealth work to the maximum benefit of the least advantaged (Rawls's difference principle).

The two principles are offered as guides for judgments about the justice of the basic structure of society, which is "the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation."[41] Why these two principles? Writing in the social contract tradition, Rawls defends them by arguing that they would be chosen by the members of society, reasoning under fair conditions. These hypothetical, fair conditions for reasoning about justice are designed to capture something fundamental to the idea of a fair society. In a fair society, how we fare over the course of our lives—our access to such fundamental goods as liberties, opportunities, and material resources——is not be fixed by the morally irrelevant differences between and among us, including differences in income and wealth, sex, race, basic values, religious convictions.[42] To capture this idea of fair social cooperation in deciding on principles of justice, Rawls proposes that people reason about the basic principles of justice for their society under conditions of ignorance about these morally irrelevant factors. The idea is to model the irrelevance that is essential to fair cooperation through ignorance of those irrelevant conditions.[43]

This hypothetical initial situation is known as the "original position", and people choose principles under what is referred to as a "veil of ignorance." Rawls calls his conception of justice "justice as fairness." The idea is that we specify fair terms of social cooperation through choice in the original position because the veil of ignorance[44] requires that we reason about principles for the basic structure while putting aside (under a veil of ignorance about) our morally irrelevant differences, and relying only on characteristics that are common to free and equal persons. It is essential to understand that this construction—choice under the "veil of ignorance"—is specific to the problem of settling on the fundamental principles for fair social cooperation among free and equal persons. It is not intended for—and there is no reason to think that it can be used for—developing fair resolutions of other social and political issues. Quite to the contrary: the point is that fair resolutions of other issues—including a just constitution and just legislation[45]—need to be guided by the principles of justice, not chosen in ignorance of them or designed using the same tools used to justify the principles.

Discussions of Rawls and algorithmic fairness have often focused on applying Rawls' conceptual devices such as the veil of ignorance and the maximin rule of choice (which Rawls discusses in connection with his original position argument) to the problem of allocating predictions (specifically prediction errors) across subgroups[46]—rather than on the substance of Rawlsian principles. While prediction errors do impose a burden on the decision subject or the user of an AI product, it is unclear whether improving an algorithm's predictions for the group that has the "least accurate predictions" is necessary or sufficient for the protections of basic liberties or for a fairer distribution of benefits and opportunities in society. The answer depends on who the members of affected group are, as well as the significance of the good that is being allocated by an algorithm in a person's life course.

Instead of using Rawlsian justice as a lens for assessing accuracy disparities in a prediction algorithm—a task at some considerable distance from its guiding purpose—we propose to begin with Rawlsian principles of justice, rooted in the idea of fair cooperation, and consider how algorithms might be implicated in helping to realize or undermine them. In the discussion that follows, we outline this approach and explore its potential to direct attention to a range of issues—about whether machine learning will help to produce a more fair society—that are not the central focus of current literature on fairness and machine learning.

4.2. We have already spoken to the issues of equal opportunity, so let's consider the **difference principle**, which requires that inequalities work to the maximum benefit of the least advantaged group: say, people living in the lowest quintile of the distribution of income and wealth. It is hard to see the inequalities that have developed in the United States over the past four decades as anything but a very stark violation of this principle.[47] The computer-based automation of the past 40 years—associated with a reduction of routine cognitive work—has arguably played a large role in shaping current inequalities.[48] Using the difference principle as a guiding principle going forward, we need to consider how the adoption of generative AI (GAI) might further shape the distribution of income and wealth, focusing in particular on the implications for the least advantaged social group.

The recent success of GAI suggests a shift in AI's ability to perform tasks that are specific to "non-routine cognitive" jobs.[49] Unlike earlier digital technologies, which primarily impacted "routine cognitive" tasks that can be captured in rules, machine learning discerns complex and implicit patterns in the way humans carry out tasks that require higher levels of cognitive engagement.[50] A few recent experimental studies suggest that large language models (LLMs) and tools powered by them, like chatbots, could potentially benefit workers with lower skill levels, reducing the skew in performance.[51] The reason for the benefits

is that LLMs are capable—through patterns they recognize in their training data—of capturing the tacit knowledge of more skilled workers (say, higher-performing customer service representatives, financial advisors, or business consultants), and transferring these insights to their less skilled colleagues.

Does this reduction in skew suggest good news for the difference principle? There is much uncertainty. (Arguments about the number of jobs that will be *impacted* by GAI are not especially helpful in reducing the uncertainty, both because "job impact" covers the gamut from task amplification to job loss, and because impact on current jobs does not tell us anything about newly created jobs.) There may be beneficial effects, if deliberate efforts are made to find ways to design and deploy GAI to augment human expertise.[52] But maybe not, especially in an environment with persistently low levels of private sector unionization, stark disparities in power between employers and employed, and a policy environment that includes tax incentives to invest in automation technologies.[53] Employers could restructure non-routine cognitive roles to reduce human employment, pushing more people into the low end of the labor market, depressing wages there. So the broader consequences could include depressed wages, reduced employment opportunities in the higher end of the labor market, and a devaluation of human labor. In short, a less fair society, measured by the standards of the difference principle (and by other standards as well).

The issue of job displacement or downgrading by GAI technologies presents a significant socio-political challenge. Auditing models for fairness—ensuring they are not biased or toxic—is very important[54], but does not begin to address this. The use of machine learning models in organizational decisions might be perfectly fair, measured by standards of organizational fairness, but the social results disastrously unfair. We are finding ourselves in another installment of the race between education and technology[55]: as AI proliferates, the education system may struggle to prepare people for the changed landscape of work. In the absence of deliberate policy measures designed to promote continuous education, re-skilling, worker empowerment, and a more equitable sharing of AI's economic gains, even the most useful and beneficent versions of GAI models—when deployed widely—might well leave us with even larger disparities between the actual circumstances in the least advantaged social group and the maximized minimum required under the difference principle.

Our comments on this issue are deliberately, and necessarily, tentative. Our aim here is not to resolve the questions raised—to provide definitive predictions and remedies—but to highlight the importance of a larger space of discussion about fairness and ML, on at least one plausible view about fairness, which requires that we give principal attention to the expectations of the least advantaged social groups.

4.3. Now consider the impact of AI and algorithms on basic liberties. We will focus on the **political liberties** associated with democracy. Democracy, as we understand it, is not only a competitive electoral system with a peaceful transfer of power, as essential as that is, but also a system of inclusive opportunities for participation and for engagement in political deliberation—public reasoning—with other citizens.[56]

There are two conditions here: *inclusion* and *deliberation*. Rawls captures inclusion in his requirement of a "fair value of political liberty," which extends an equal opportunity requirement to the democratic process.[57] Political liberty has its fair value when equally motivated and equally able citizens have equal chances for political influence. Deliberativeness, on the other hand, is about the quality of public reflection and argument; it requires a public sphere where common issues are debated and the terms of disagreement clarified based on shared information and reasons.[58] This requirement is captured in Rawls's conception of "public reason."[59] Together, these conditions require a political environment where individuals with diverse values and beliefs engage in the exchange of ideas and arguments, and where this exchange is conducted on an equal and inclusive basis to shape the laws and policies they will live under.

While BHN mention concerns around unfairness in ad targeting—including political ads[60]—their critique focuses on group disparities in the types of ads delivered (22). Under a broader understanding of fair political liberties, the central concern is not the disparate treatment of groups, but the potential reduction in shared information, a common space of discussion, and deliberativeness as a result of hyper-personalized content targeting on algorithmically-powered social media platforms. AI—by enabling fine-grained inferences about individuals from patterns in massive amounts of user data—has arguably created new economic incentives[61] for fragmenting public discourse.

The shift in perspective from organizational fairness to an expansive view of a fair value for political liberty is important as well for thinking of solutions to the problems we have named. Enhancing deliberativeness in online discussion is not primarily a matter of reducing group disparities in say, the targeting of political ads. It might require a deeper sense of responsibility on the part of contributors to public discussion who have taken on the role of authors but not always with a sense about the responsibilities

of authorship, a more thoughtful understanding of the extent to which current pathologies of political debate come from political discussion on platforms[62], improved methods for fostering digital literacy, as well as socio-technical innovation—beyond engineering solutions such as Meta's Variance Reduction System.[63] These solutions look beyond organizational decision making, but many require the active engagement of organizations in service of broader sociopolitical concerns.

Consider also the use of GAI for political influence. Here again, there are issues (both challenges and opportunities) about inclusivity and deliberativeness. The deliberate production of compellingly persuasive but untruthful information under the proliferation of AI sources may impinge on the opportunity to exercise informed influence as a citizen.[64] This may be accomplished via deliberate deceit, using deepfakes and other AI-generated content[65], though misinformation may also result from hallucinations.

A recent study by Costello, Pennycook, and Rand examined the potential of large language models (LLMs) to *reduce* belief in conspiracy theories.[66] In a three-round dialogue, participants were asked to state a conspiracy theory they believed in and to explain the evidence and arguments for their belief. GPT-4 Turbo was prompted to engage in a discussion that focused on debunking these beliefs, responding directly to the participants' reasons. The results showed that this personalized engagement reduced participants' belief in the conspiracy theory by roughly 20%, across a range of pre-existing confidence in belief. Of course, you might wonder if Turbo would be equally effective in persuading people to believe in a conspiracy theory, if prompted to do so. That is an open question. But still, these findings underscore that we need to be thinking not simply about democratic risks but also about possibilities of *improving* deliberativeness.

Finally, one might also worry, in connection with the fair value of political liberties, about asymmetries in the access to such technologies of influence. GAI has the potential to target a particular political message to the individual; that is effective political influence at enormous scale. But how will access to these automated, personalizable loud speakers be settled? State of the art language models currently contain hundreds of billions of parameters and are expensive to train and fine-tune.[67] Public access to the parameters of these models as well as the data around their usage may be extremely limited (depending on what happens with open source models). Companies developing these proprietary tools have exclusive control over the types of political content and perspectives that their models may produce—via deliberate optimization of model outputs—while average citizens have effectively no say in such moderations.

None of the concerns we have sketched are primarily about the algorithms themselves. But they are genuine concerns about fairness and machine learning. By starting with a conception of a just society, with fair terms of cooperation, our attention is drawn outside the decisions of organizations and into this larger space of issues.

5 CONCLUSION

FaML is a rich and complex contribution to the literature on ML and fairness that powerfully merges normative thought with the formal aspects of machine learning.

While it touches on a wide range of topics, it focuses principally on organizational decisions, in particular on correcting historical group disadvantage. This focus is characteristic of the scholarship on the topic thus far, especially in the United States. We have discussed three limitations of their approach: discrimination beyond group subordination, equality of opportunity beyond organizational decisions, and justice beyond the equality of opportunity. Ultimately, in considering what BHN have left out of their main discussion, we find ourselves reaching for a more forward-looking conception of fairness and machine learning that integrates political theory and is attentive to broader sociopolitical realities.

Consider legislative efforts beyond the US context. The EU's AI Act serves as a pertinent example of how regulatory frameworks can integrate broader concerns for equality, human rights, and democracy with the technical sophistication in formalization, measurement, and oversight. As the emerging regulatory landscape in the United States evolves, there is a great need for all disciplines to contribute to AI policy discourse. While recognizing the great importance of the topics discussed, we hope the points raised in this essay will help to widen the scope of current discussions and further the multidisciplinary work that Barocas, Hardt and Narayanan have helped pioneer—individually and collectively—in FaML.

[1] This essay, appearing in the *ACM Journal of Responsible Computing,* discusses Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (Cambridge, MA: MIT Press, 2023). All references to the book are included parenthetically in the text. We are grateful for comments on an earlier draft from Linda Eggert, Archon Fung, Johannes Himmelreich, Norman Mu, Rob Reich, David Robinson, Thomas Scanlon, and Kara Schechtman.

[2] The expression "machine executed algorithm" may strike some as pleonastic. Unfortunately, much public discussion seems to overlook the fact that a person or a collection of people can execute an algorithm. An algorithm is an algorithm whether it is executed by silicon units or carbon units. The fact that a human being is involved in a recommendation or decision does not imply that the recommendation or decision is not algorithmic.

[3] Annex 3 on high-risk categories in the EU's AI Act provides a pretty good list of consequential uses specifically of AI. See https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[4] http://gendershades.org/

[5] Eugene Volokh, "Large Libel Models: Liability for AI Output," https://www.journaloffreespeechlaw.org/volokh4.pdf.

[6] FaML is self-consciously US-centric (xiii). Chap. 6 explains the basics of US anti-discrimination law, with no discussion of, for example, the 2000 EU Equality Directives, or the equality principle and anti-discrimination provisions of the Charter of Fundamental Rights, legally binding since 2009. For discussion, see *EU Anti-Discrimination Law Beyond Gender*, eds. Uladzislau Belavusau and Kristin Henrard (Bloomsbury Publishing, 2023); Sara Tommasi, *The Risk of Discrimination in the Digital Market: From the Digital Services Act to the Future* (Springer: Cham, 2023).

[7] See for example, FaML, and references therein.

[8] They treat "*unfairness* and *discrimination* roughly synonymously" (83, 186).

[9] On the relationship between discrimination, classification, group-disadvantaging, and group subordination, see Owen Fiss, "Another Equality," *Issues in Legal Scholarship* (Berkeley Electronic Press, 2004); Jack M. Balkin and Reva B. Siegel, "The American Civil Rights Tradition: Anticlassification or Antisubordination?," *University of Miami Law Review 58, 9* (2003), pp. 9-33.

[10] Earlier work such as Binns (2018) explored the conceptual foundations of algorithmic fairness from the perspective of moral and political philosophy.
Binns, Reuben. "Fairness in machine learning: Lessons from political philosophy." *Conference on fairness, accountability and transparency*. PMLR, 2018.

[11] Hedden, Brian. "On Statistical Criteria of Algorithmic Fairness." Philosophy & Public Affairs, vol. 49, no. 2, 2021, pp. 209-231; Green, Ben. "The Flaws of Policies Requiring Human Oversight of Government Algorithms." Computer, vol. 55, no. 10, 2022, pp. 55-65; Fazelpour, Sina, and Zachary C. Lipton. "Algorithmic Fairness from a Non-ideal Perspective." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 57-63; Fazelpour, Sina, Zachary C. Lipton, and David Danks. "Algorithmic fairness and the situated dynamics of justice." *Canadian Journal of Philosophy* 52.1 (2022): 44-60.

[12] Liu, Lydia T., et al. "Delayed Impact of Fair Machine Learning." Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018, pp. 3150-3158; Corbett-Davies, Sam, et al. "The measure and mismeasure of fairness." *The Journal of Machine Learning Research* 24.1 (2023): 14730-14846.

[13] Key contributions to this earlier literature include: Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. 2012. Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big data* 5.2 (2017): 153-163. Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." 8th Innovations in Theoretical Computer Science Conference (ITCS 2017).

[14] For instance, Herbert-Johnson et al (2018) began a line of work based on the notion of algorithmic fairness that "guarantees meaningful (calibrated) predictions for every sub-population that can be identified within a specified class of computations." Similarly, Kearns et al (2018) suggested that one "would like to be able to satisfy a fairness constraint […] for a combinatorially large or even infinite collection of structured subgroups definable over protected attributes." Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, Guy N. Rothblum, "Multicalibration: Calibration for the (computationally-identifiable) masses." *International Conference on Machine Learning*. PMLR, 2018. Kearns, Michael, et al. "An empirical study of rich subgroup fairness for machine learning." *Proceedings of the conference on fairness, accountability, and transparency*. 2019.

[15] Though see FaML, 26-29 for concerns about how using machine learning might displace existing procedural protections against bureaucratic arbitrariness.

[16] Recall the terminological point: "we use the terms *unfairness* and *discrimination* roughly synonymously" (83).

[17] BHN offer John Rawls' conception of fair equality of opportunity (see below, n. 29) as an example of what they call the "broad view of equality of opportunity." The broad view, as they say, is "not really about fairness in decision making; it is about the design of society's basic institutions. See pp. 90-91.

[18] Of course there are very good reasons for doing so.

[19]" The most fundamental idea in this conception of justice is the idea of society as a fair system of social cooperation over time from one generation to the next." John Rawls, *Justice as Fairness: A Restatement*, ed. Erin Kelly (Cambridge, MA: Harvard University Press, 2001), p. 5.

[20] Ibid., pp. 18-24.

[21] See Johannes Himmelreich, "Ethics of technology needs more political philosophy," *Communications of the ACM* 63, 1 (2019): 33–35; Gabriel, Iason. "Toward a theory of justice for artificial intelligence."

Daedalus 151.2 (2022): 218-231; Claire Benn and Seth Lazar, "What's wrong with automated influence," *Canadian Journal of Philosophy* 52, 1 (2022): 125-148.

[22] On tensions between the anti-classification view of equal protection and fairness through awareness, see Daniel E. Ho and Alice Xiang, "Affirmative Algorithms: The Legal Grounds for Fairness as Awareness," The University of Chicago Law Review Online Archive, https://lawreviewblog.uchicago.edu/2020/10/30/aa-ho-xiang/.

[23] We focus on the US employment context, but our point would remain in the EU setting, guided by the Employment Equality Directive. https://ec.europa.eu/commission/presscorner/detail/en/MEMO_08_69

[24] Groff v. DeJoy, 600 U.S. ___ (2023).

[25] Trans World Airlines, Inc. v. Hardison, 432 U.S. 63 (1977), https://supreme.justia.com/cases/federal/us/432/63/

[26] Gerald Groff, the petitioner in Groff v. DeJoy, is an evangelical Christian who regards Sunday not only as a day of worship, but as a day of rest.

[27] The classic statement of the notion of non-comparative justice and fairness is in Joel Feinberg, "Noncomparative Justice," *Philosophical Review*, 83, pp. 297–338. For discussion in the context of algorithmic fairness, see Deborah Hellman, "Algorithmic Fairness," unpublished draft.

[28] Here we are not offering an account of what a group is, or which kinds of group disadvantaging are matters of concern; we are merely stating that the claimant's concern is not about groups.

[29] Eugene Volokh, "Should the Law Limit Private-Employer-Imposed Speech Restrictions," *Journal of Free Speech Law* 2, 1 (2022): 269-298.

[30] Joseph Fishkin, *Bottlenecks* and "The Anti-Bottleneck Principle in Employment Discrimination Law," *Washington University Law Review* 91 (2014).

[31] We are not claiming that they had the desired effect. On ban the box laws, see Amanda Agan and Sonja Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *The Quarterly Journal of Economics*, 133, 1 (February 2018): 191–235..

[32] This is a rough statement of Rawls's idea of fair equality of opportunity, which BHN endorse in their discussion of the broad view (91). See John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 2002), pp. 73-78.

[33] David Robinson, *Voices in the Code: A Story about People, Their Values, and the Algorithm They Made* (New York: Russell Sage Foundation, 2022).

[34] See, for example, the algorithm developed by the University of Pittsburgh critical care department. (https://bioethics.pitt.edu/sites/default/files/Univ Pittsburgh - Allocation of Scarce Critical Care Resources During a Public Health Emergency.pdf)

[35] https://inside.upmc.com/how-should-we-ethically-allocate-scare-critical-care-resources-during-covid-19-pandemic/; https://www.bumc.bu.edu/gimcovid/files/2020/04/Pittsburgh-Model-hospital-policy-for-allocation-of-critical-care_2020-03-23.pdf.

[36] Michael Marmot, *The Health Gap* (London: Bloomsbury Press, 2015).

[37] For discussion of some related puzzles, see Lily Hu and Issa Kohler-Hausmann. "What's sex got to do with machine learning?." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 513-513. 2020.

[38] Robert Manduca and Robert J. Sampson. "Punishing and toxic neighborhood environments independently predict the intergenerational social mobility of black and white children." *Proceedings of the national academy of sciences* 116.16 (2019): 7772-7777. https://www.pnas.org/doi/10.1073/pnas.1820464116

[39] Hickey, Colin, et al. "The agents of justice." *Philosophy Compass* 16.10 (2021): e12770.

[40] John Rawls, *Political Liberalism*, expanded edition (New York: Columbia University Press, 2005), pp. 5-6.

[41] Rawls, *Theory*, p. 6. There is a large literature on what the basic structure is and whether the idea of developing principles specifically for the basic structure is the right approach for a theory of justice. See in particular G.A. Cohen, "Where the Action Is: On the Site of Distributive Justice," *Philosophy and Public Affairs* 26 (1997): 3-30.

[42] BHN note the connection between fairness and independence of irrelevant considerations at 85.

[43] Rawls, *Justice as Fairness: A* Restatement, pp. 14-18. On the rationale of the original position and how it models the distinction between relevant and irrelevant considerations, see Joshua Cohen, "The Original Position and Scanlon's Contractualism," in *The Original Position*, ed. Timothy Hinton (Cambridge: Cambridge University Press, 2015), pp. 179-200.

[44] Together with a few other conditions, including the ideas of primary goods and mutually disinterested rationality.

[45] See Rawls on the four-stage sequence for applying the two principles of justice. Rawls, *Theory*, pp. 171-76.

[46] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang, "Fairness without demographics in repeated loss minimization." In *International Conference on Machine Learning*, pp. 1929-1938. PMLR, 2018; Heidari, Hoda, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. "Fairness behind a veil of ignorance: a welfare analysis for automated decision making." *Advances in Neural Information Processing Systems 31* (2019): 1265-1276; Franke, Ulrik. "Rawls's original position and algorithmic fairness." *Philosophy & Technology* 34.4 (2021): 1803-1817; Barsotti, Flavia, and Rüya Gökhan Koçer. "MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias." *AI & SOCIETY* (2022): 1-14.

[47] Though for some qualification, see David Autor, Arindrajit Dube, and Annie McGrew. 2023. "The Unexpected Compression: Competition at Work in the Low Wage Labor Market." National Bureau of Economic Research Working Paper no. 31010, March 2023.

[48] See Daron Acemoglu and Pascual Restrepo, "Tasks, Automation, and the Rise in US Wage Inequality," *Econometrica* 90, 5 (September 2022): 1973-2016; David Autor, Caroline Chia, Anna Salomon's, Bryan Seegmiller, *New Frontiers: The Origins and Content of New Work, 1940–2018.* Other factors that bear

particularly on the difference principle include the continuing decline of private sector unions and the impact of hyper-globalization on manufacturing employment. [China Shock]

[49] David Autor,"The labor market impacts of technological change: from unbridled enthusiasm to qualified optimism to vast uncertainty," NBER Working Paper 30074 http://www.nber.org/papers/w30074.

[50] Tyna Eloundou, et al. "Gpts are gpts: An early look at the labor market impact potential of large language models." *arXiv preprint arXiv:2303.10130* (2023).

[51] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond, "Generative AI at work," No. w31161. National Bureau of Economic Research, 2023. [Also the BCG study, and other refs] Candelon, F. et al. (2023) How people can create-and destroy-value with Generative AI, BCG Global. Available at: https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai (Accessed: 16 April 2024);

Shakked Noy and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," Working Paper, March 2, 2023.

[52] David Autor, "AI Could Actually Help Rebuild The Middle Class," *Noema* February 12, 2024. https://www.noemamag.com/how-ai-could-help-rebuild-the-middle-class/

[53] See, for example, Daron Acemoglu, https://www.congress.gov/117/meeting/house/114205/witnesses/HHRG-117-EF00-Wstate-AcemogluK-20211103.pdf

[54] See, for example, Liang, Percy, et al. "Holistic Evaluation of Language Models." Transactions on Machine Learning Research (2023).

[55] Claudia Goldin and Lawrence F. Katz, *The race between education and technology* (Cambridge, MA: Harvard University Press, 2009).

[56] The idea that democracy is characterized by a competitive electoral system with a peaceful transfer of power is sometimes called a "minimalist" conception of democracy; see Joseph Schumpeter, *Capitalism, Socialism, and Democracy* (1942) and Adam Przeworski, *Why Bother with Elections?* (2018).

57 Rawls, Theory, pp. 224-228, 233-234; Political Liberalism, pp. 327-331.

[58] On deliberation and clarifying the terms of agreement, see Simon Niemeyer, Francesco Veri, John Dryzek, Andre Bachtiger, "How Deliberation Happens: Enabling Deliberative Reason," *American Political Science Review* (2023): 1–18. More generally, see John Dryzek, et al., "The crisis of democracy and the science of deliberation," Science, 363, 6432 (March 2019): 1144-1146. https://www.science.org/doi/10.1126/science.aaw2694

[59] Rawls, *Political Liberalism*, Lecture 6.

[60] Ali, Muhammad, et al. "Ad delivery algorithms: The hidden arbiters of political messaging." *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 2021.

[61] Benn and Lazar, "What's wrong with automated influence."

[62] For a skeptical assessment of folk wisdom, see, for example, Andrew Guess, et al., "How do social media feed algorithms affect attitudes and behavior in an election campaign?," *Science* 381 (2023): 398–404.

[63] See Joshua Cohen and Archon Fung, "Democracy and the Digital Public Sphere, " in *Digital Technology and Democratic Theory, eds. Lucy Bernholz, Helene Landemore, and Rob Reich* (Chicago: University of Chicago Press, 2021), pp. 23-61, and "Democratic responsibility in the digital public sphere," *Constellations* 30 (2023) :92–97. On the Meta system, see https://ai.meta.com/blog/advertising-fairness-variance-reduction-system-vrs/.

[64] On persuasiveness, see Esin Durmus and Liane Lovitt and Alex Tamkin and Stuart Ritchie and Jack Clark, and Deep Ganguli, "Measuring the Persuasiveness of Language Models, April 9, 2024, https://www.anthropic.com/news/measuring-model-persuasiveness; on persuasive propaganda, see Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, Michael Tomz, "How persuasive is AI-generated propaganda?," *PNAS Nexus*, 3, 2 (February 2024): 1-7.

[65] Valerie Wirtschafter, "The impact of generative AI in a global election year." (2024).

[66] Thomas H. Costello, Gordon Pennycook, David G. Rand, "Durably reducing conspiracy beliefs through dialogues with AI," *Science* 385, eadq1814 (2024).

[67] See, for example, Table 5 in Raiaan, Mohaimenul Azam Khan, et al. "A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges." *IEEE Access* (2024).