7

A broader view of discrimination

Machine learning systems don't operate in a vacuum; they are adopted in societies that already have many types of discrimination intertwined with systems of oppression such as racism. This is at the root of fairness concerns in machine learning. In this chapter we'll take a systematic look at discrimination in society. This will give us a more complete picture of the potential harmful impacts of machine learning. We will see that while a wide variety of fairness interventions are possible—and necessary—only a small fraction of them translate to technical fixes.

Case study: the gender earnings gap on Uber

We'll use a paper that analyzes the gender earnings gap on Uber[1] as a way to apply some of the lessons from the previous two chapters while setting up some of the themes of this chapter.[1] The paper starts with the observation that female drivers earn 7% less on Uber per active hour than male drivers do. It concludes that this gap can be explained by three factors: gender differences in drivers' choices of where to drive, men's greater experience on the platform, and men's tendency to drive faster. It finds that customer discrimination and algorithmic discrimination do not contribute to the gap. We'll take the paper's technical claims at face value, but use the critical framework we've introduced to interpret the findings quite differently from the authors.

First, let's understand the findings in more detail.

The paper analyzes observational data on trips in the United States, primarily in Chicago. Above, we've drawn a causal graph showing what we consider to be the core of the causal model studied in the paper (the authors do not draw such a graph and do not pose their questions in a causal framework; we have chosen to do so for pedagogical purposes). A full graph would be much larger than the Figure; for example, we've omitted a number of additional controls, such as race, that are presented in the appendices.

We'll use this graph to describe the findings. At a high level, the graph describes a joint distribution whose samples are trips. To illustrate, different trips corresponding to the same driver will have the same *Residence* (unless the driver moved during their tenure on the platform), but different *Experience* (measured as number of prior trips).

[1] The study was coauthored by current and former Uber employees.

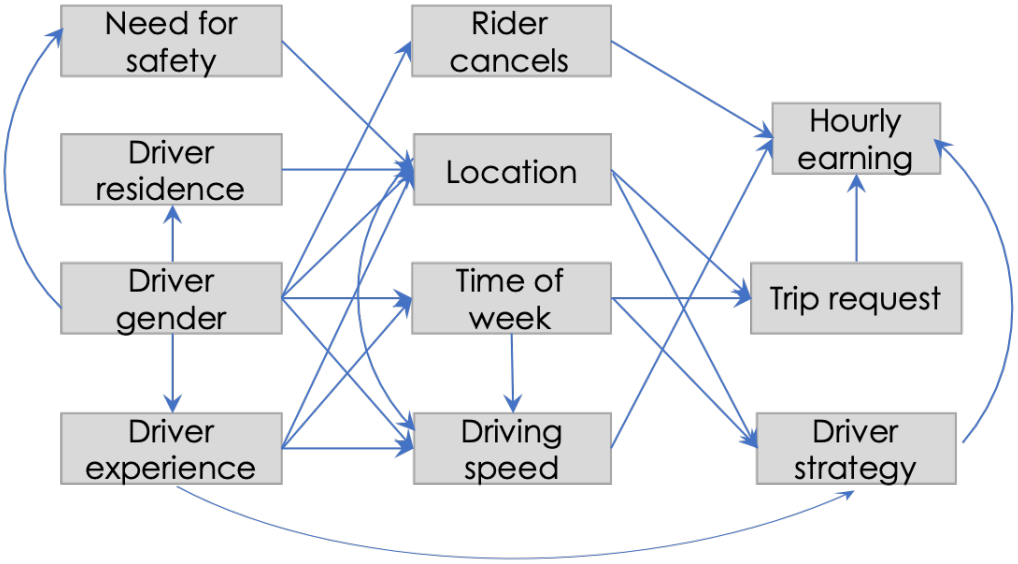

Figure 1: Our understanding of the causal model implicit in the Uber study.

Drivers' hourly earnings are primarily determined by the algorithm that allocates trip requests from riders to drivers. The allocation depends on demand, which in turn varies by location and time of the week (the week-to-week variation is considered noise). Uber's algorithm ignores driver attributes including experience and gender, hence there are no arrows from those nodes to *Trip request*. In addition, a few other factors might affect earnings. Drivers who drive faster complete more trips, drivers may strategically accept or cancel trips, and riders might discriminate by cancelling trips after the driver accepts.

The paper uses a technique called Gelbach decomposition[2] to identify the effect of each of several variables on the hourly earnings. It finds that the earnings gap (i.e. effect of *Driver gender* on *Hourly earning*) can be entirely explained by paths involving *Driver experience*, *Location*, and *Driving speed*. Paths through *Rider cancellation* and *Time of week* don't have significant effects.

The authors further interrogate the effect of gender on location (i.e. the choice of where to drive), and find that women are less likely to drive in less safe areas that also turn out to be more lucrative. They then dig deeper and argue that this effect operates almost entirely by women *residing* in safer areas and choosing to drive based on where they live.

The returns to experience could operate in several ways. The authors don't decompose the effect but suggest several possibilities: the choice of where and when to drive and other elements of strategy including which rides to accept. A

[2]Decomposition is a set of techniques used in economics for quantifying the contribution of various sources to an observed difference in outcomes. Although the authors don't perform causal inference, we will continue to talk about their findings in causal terms for pedagogical purposes. The difference is not salient to the high-level points we wish to make.

key finding of the paper is the effect of gender on experience. Men are less likely to leave the platform and drive more hours during each week that they stay on the platform, resulting in a large experience differential. There are no gender differences in *learning* from experience: male and female drivers' behavior changes at the same rate for a given number of trips.

The paper highlights questions that can be studied using observational data but not necessarily with field experiments (audit studies). An audit study of the Uber gender pay gap (along the lines of those discussed in the previous chapter) may have involved varying the driver's name to test the effect on rider cancellation and ratings. Such an experiment would have no way to uncover the numerous other paths by which gender affects earnings.[3]

Causal diagrams in realistic scenarios are more complex than typical textbook examples. We reiterate that the graph above is much simplified compared to the (implicit) graph in the paper. The estimation in the paper proceeds as a series of regressions focusing iteratively on small parts of the graph, rather than an analysis of the entire graph at once. In any messy exercise such as this, there is always the possibility of unobserved confounders.

Despite the number of possible effects considered in the study, it leaves out many others. For example, some drivers may move to take advantage of the earning potential. This would introduce a cycle into our causal graph (*Location –> Residence*). This type of behavior might seem unlikely for an individual driver, which justifies ignoring such effects in the analysis. Over time, however, the introduction of transportation systems has the potential to reshape communities.[3,4] Today's empirical methods have limitations in understanding these types of long-term phenomena that involve feedback loops.

A more notable omission from the paper is the effect of driver gender on experience. Why do women drop off the platform far more frequently? Could one reason be that they face more harassment from riders? The authors don't seem to consider this question.

This leads to our most salient observation about this study: the narrow definition of discrimination. First, as noted, the study doesn't consider that differential dropout rates might be due to discrimination.[4] This is especially pertinent since the gender gap in hourly earnings is merely 7% whereas the gap in participation rate is a factor of 2.7! One would think that if there is rider discrimination, it would be most apparent in its effect on dropout rates. In contrast, the only avenue of discrimination considered in the paper involves a (presumably misogynistic) rider who cancels a ride, incurring delays and potentially algorithmic penalties, based solely on the driver's gender.

Further, the authors take an essentialist view of the gender difference in average

[3] An audit study would be more suited for studying discrimination *by drivers against riders*, in part because drivers in these systems exercise more choice in the matching process than riders do. Indeed, a study found that UberX and Lyft drivers discriminate against Black and female riders.[2]

[4] For example, the authors say in the abstract: *Our results suggest that, in a "gig" economy setting with no gender discrimination and highly flexible labor markets, women's relatively high opportunity cost of non-paid-work time and gender-based differences in preferences and constraints can sustain a gender pay gap.*

speed (e.g. "men are more risk tolerant and aggressive than women"). We may question how innate these differences are, given that in contemporary U.S. society, women may face social penalties when they are perceived as aggressive. If this is true of driver-rider interactions, then women who drive as fast as men will receive lower ratings with attendant negative consequences. This is a form of discrimination by riders. [5]

Another possible view of the speed difference, also not considered by the authors, is that male drivers on average provide a lower quality of service due to an increase in accident risk resulting from greater speed (which also creates negative externalities for others on the road). In this view, Uber's matching algorithm discriminates against female drivers by *not* accounting for this difference.[6]

Finally, the paper doesn't consider structural discrimination. It finds that women reside in less lucrative neighborhoods and that their driving behavior is shaped by safety considerations. However, a deeper understanding the reasons for these differences is outside the scope of the paper. In fact, gender differences in safety risks and the affordability of residential neighborhoods can be seen as an example of the greater burden that society places on women. In other words, Uber operates in a society in which women face discrimation and have unequal access to opportunities, and the platform perpetuates those differences in the form of a pay gap.[7]

Let us generalize a bit. There is a large set of studies that seek to explain the reasons for observed disparities in wages or another outcome. Generally these studies find that the direct effect of gender, race, or another sensitive attribute is much smaller than the indirect effect. Frequently this leads to a vigorous debate on whether or not the findings constitute evidence of discrimination or unfairness. There is room for different views on this question. The authors of the Uber study interpreted none of the three paths by which gender impacts earnings—experience, speed, and location—as discrimination; we've argued that all three can plausibly be interpreted as discrimination. Different moral frameworks will lead to different answers. Views on these questions are also politically split. As well, scholars in different fields often tend to answer these questions differently (including, famously, social science and economics[7]).

Certainly these definitional questions are important. However, perhaps the greatest value of studies on mechanisms of discrimination is that they suggest avenues for intervention *without* having to resolve definitional questions. Looking at the Uber study from this lens, several interventions are apparent. Recall that there is a massive gender disparity in the rate at which drivers drop out of the platform. Uber could more actively solicit and listen to feedback from female drivers and use that feedback to inform the design of the app. This may lead to

[5]For a broad discussion of customer ratings as a vehicle for discrimination on Uber, see.[5]

[6]If riders give lower ratings to drivers who drive faster at the expense of safety, then the matching algorithm does indirectly take safety considerations into account. We think it is unlikely that driver ratings adequately reflect the risks of speeding, due to cognitive biases. After all, that is why we need speed limits instead of leaving it up to drivers.

[7]See[6] for a discussion of many ways in which existing geographic inequalities manifest in sharing economy platforms including Uber.

interventions such as making it easier for drivers (and riders) to report harassment and taking stronger action in response to such reports.

As for the speed difference, Uber could warn drivers who exceed the speed limit or whose speed results in a predicted accident risk that crosses some threshold (such a prediction is presumably possible given Uber's access to data). In addition, Uber could use its predictive tools to educate drivers about strategy, decreasing the returns to experience for all drivers. Finally, the findings also give greater urgency to structural efforts to make neighborhoods safe for women. None of these interventions require a consensus on whether or not female drivers on Uber are discriminated against.

Three levels of discrimination

Sociologists organize discrimination into three levels: structural, organizational, and interpersonal.[8,7] Structural discrimination arises from the ways in which society is organized, both through relatively hard constraints such as discriminatory laws and through softer ones such as norms and customs. Organizational factors operate at the level of organizations or other decision-making units, such as a company making hiring decisions. Interpersonal factors refer to the attitudes and beliefs that result in discriminatory behavior by individuals.

A separate way to classify discrimination is as direct or indirect. By direct discrimination we mean actions or decision processes that make explicit reference to a sensitive attribute. By indirect discrimination we refer to actions or decision processes that make no such reference, yet disadvantage one or more groups. The line between direct and indirect discrimination is hazy and it is better to think of it as a spectrum rather than a binary category.[8]

Table 1: Examples of discrimination organized into three levels and on a spectrum of directness

Level	More direct	More indirect
Structural	Laws against same-sex marriage	Better funded schools in wealthier, more segregated areas
Organizational	Lack of disability accommodations	Networked hiring
Interpersonal	Overt animus	Belief in need for innate brilliance (combined with gender stereotypes)

[8]For attempts by philosophers to formalize the distinction, see.[9] For a technical treatment of direct vs. indirect effects, refer back to the Causality chapter. See also;[10] in particular, the point that "any direct effect is really an indirect effect if you zoom further into the relevant causal mechanism".

Structural factors

Structural factors refer to ways in which society is organized. A law that overtly limits opportunities for certain groups is an example of a direct structural factor. Due to various rights revolutions around the world, there are fewer of these laws today than there used to be. Yet, discriminatory laws are far from a thing of the past. For example, as of 2021, a mere 29 countries recognize marriage equality.[11] Further, discriminatory laws of the past have created structural effects which persist today.[12]

Indirect structural discrimination is pervasive in virtually every society. Here are two well known examples affecting the United States. Drug laws and drug policies, despite being facially neutral, have the effect of disproportionately affecting minority groups, especially Black people.[13] Schools in high-income neighborhoods tend to be better funded (since public schools are funded primarily through property taxes) and attract more qualified teachers, transmitting an educational advantage to children of higher-income parents.

Other factors are even less tangible yet no less serious in terms of their effects, such as cultural norms and stereotypes. In the case study of gender bias in Berkeley graduate admissions in Chapter 5, we encountered the hypothesis that societal stereotypes influence people's career choices in a way that reproduces gender inequalities in income and status:

> The bias in the aggregated data stems ... apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

Organizational factors

Organizational factors operate at the level of organizations or decision-making units: how they are structured, the decision making rules and processes they put in place, and the context in which individual actors operate. Again, these lie on a spectrum between direct and indirect.

The most direct form of discrimination—excluding people from participation explicitly based on group membership—is mostly unlawful in liberal democracies. However, practices such as lack of disability accommodations and failure to combat sexual harassment are rampant. A more indirectly discriminatory policy is the use of employees' social networks in hiring, an extremely common practice. One observational study found that the use of employee referrals in predominantly White firms reduced the probability of a Black hire by nearly 75% relative to the use of newspaper ads.[14] The study controlled for spatial segregation, occupational segregation, city, and firm size.

Organizational discrimination can be revealed and addressed at the level of a

single organization, unlike structural factors (e.g. no individual school is responsible for teachers being attracted to schools in high-income neighborhoods).

Interpersonal factors

Interpersonal factors refer to the attitudes and beliefs that result in discriminatory behavior by individuals. Sometimes people may discriminate because of an overt animus for a certain group, in the sense that the discriminator does not attempt to justify it by any appeal to rationality.

More often, the mechanisms involved are relatively indirect. A 2015 study found that academic fields in which achievement is believed to be driven by innate brilliance exhibit a greater gender disparity, i.e., they have fewer women.[15] The authors propose that the disparity is caused by the combination of the belief in the importance of innate brilliance together with stereotypes about lower innate brilliance in women. This combination could then impact women in brilliance-emphasizing disciplines in two ways: either by practitioners of those disciplines exhibiting biases against women, or by women internalizing those stereotypes and self-selecting out of those disciplines (or performing more poorly than they otherwise would). The authors don't design tests to distinguish between these competing mechanisms. However, they do test whether the observed disparities could alternatively be caused by actual innate differences (rather than beliefs in innate differences) in ability or aptitude, or willingness to work long hours. Using various proxies (such as GRE score for innate ability), they argue that such competing explanations cannot account for the observed differences.

One may wonder: can we not test for innate differences more rigorously, such as by examining young children? A follow-up study showed that children as young as six tend to internalize gendered stereotypes about innate brilliance, and these stereotypes influence their selection of activities.[16] These difficulties hint at the underlying complexity of the concept of gender, which is produced and reinforced in part through these very stereotypes.[17]

To recap, we've discussed structural, organizational, and interpersonal discrimination, and the fact that these are often indirect and pervasive. The three levels are interconnected: for example, in the Uber case study, structural inequalities don't perpetuate themselves, but rather through organizational decisions; those decisions at Uber are made by individuals whose worldviews are shaped by culture. In other words, even structural discrimination is actively perpetuated, and we collectively have the power to mitigate it and to reverse course. It would be a mistake to resign ourselves to viewing structural discrimination as simply the way the world is.

Notice that adopting statistical decision making is not automatically a way out of any of these factors, which operate for the most part in the background and not at a single, discrete moment of decision making.

The persistence and magnitude of inequality

Formal equality under the law primarily addresses direct discrimination and has relatively little effect on indirect discrimination, whether structural, organizational, or interpersonal. This is one reason why inequality can be persistent in societies that seemingly promise equal opportunity. Here are two stark examples of how long inequalities can sustain themselves.

Beginning in 1609, Jesuit missions were established in the Guaraní region of South America that overlaps modern day Argentina, Paraguay, and Brazil. In addition to religious conversion, the missionaries undertook educational efforts among the indigenous people. However, due to political upheaval in Spain and Portugal, the missions abruptly ended in 1767-68 and the missionaries were expelled. How long after this date would we expect the geographic inequalities introduced by Jesuit presence to persist? Perhaps a generation or two? Remarkably, the Jesuit effect on educational attainment has been found to persist 250 years later: areas closer to a former Mission have 10-15% higher literacy rates as well as 10% higher incomes [9]. Another study of the long-run persistence of inequality shows the present-day effects of a system of colonial forced labor in Peru in Bolivia between 1573 and 1812.[19]

More evidence for the long-run persistence of inequality comes from the city of Florence, based on a unique dataset containing tax-related data for all individuals from the year 1427. A working paper finds that surnames associated with wealthier individuals in the dataset are associated with wealthier individuals today, six hundred years later.[20]

While these are just a few examples, research shows that persistence of inequality over generations along social and geographic lines is the norm. Yet it is not widely appreciated. For example, Americans believe that an individual born into the bottom quintile of the income distribution has a 1-in-6 chance of rising to the top quintile but the observed likelihood is 1-in-20.[21] Mobility in the U.S. has decreased since the 1980s, and is lower for Black Americans than White Americans.[22]

These inequalities are significant because of their magnitude in addition to their persistence. Median income of Black Americans is about 65% that of White Americans.[23] Wealth inequality is much more severe: the median wealth of Black households is about 11% that of White households. Turning to gender, full-time, year-round working women earned 80% of what their male counterparts earned.[24] Geographic inequalities also exist. For example, the richest and poorest census tracts in the United States differ in average income by a factor of about 30.[25]

[9] See.[18] The study makes use of a clever idea to argue that the mission locations were essentially random, making this a natural experiment, and includes various checks to rule out alternative explanations.

Machine learning and structural discrimination

For a book about machine learning, we've covered a lot of ground on discrimination and inequality in society. There's a reason. To understand fairness, it isn't enough to think about the moment of decision making. We also need to ask: what impact does the adoption of machine learning by decision makers have in long-lasting cycles of structural inequality in society? Does it help us make progress toward enabling equality of opportunity, or other normative ideals, over the course of people's lives? Here are some observations that can help answer those questions.

Predictive systems tend to preserve structural advantages and disadvantages

Predictive systems tend to operate within existing institutions. When such institutions perpetuate inequality due to structural factors, predictive systems will only reify those effects, absent explicit intervention. Predictive systems tend to inherit structural discrimination because the objective functions used in predictive models usually reflect the incentives of the organizations deploying them. As an example, consider a 2019 study found strong racial bias in a system used to identify patients with a high risk of adverse health outcomes, in the sense that Black patients were assigned lower scores compared to equally at-risk White patients.[26] The authors found that this happened because the model was designed to predict healthcare costs instead of needs, and the healthcare system spends less caring for Black patients than White patients even when they have the same health conditions.

Suppose a firm makes hiring decisions based on a model that predicts job performance based on educational attainment. Imagine a society where students from higher-income families, on average, have had better educational opportunities that translate to greater job skills. This is not a measurement bias in the data that can be corrected away: education level genuinely predicts job performance. Thus, an accurate predictive system will rank higher-income candidates higher on average.

The structural effect of such systems become clear when we imagine every employer applying similar considerations. Candidates with greater educational opportunities end up with more desirable jobs and higher incomes. In other words, predictive systems have the effect of transferring advantages from one phase of life to the next, and one generation to the next.

This phenomenon shows up in less obvious ways. For instance, online ad targeting is based on the assumption that differences in past behavior between users reflect differences in preferences. But they might also result from differences in structural *circumstances*, and there is no way for targeting engines to tell the difference. This helps explain why ads, including job ads, may be targeted in ways that reinforce stereotypes and structural discrimination. [10]

This aspect of predictive systems is amplified by compounding injustice.[28,29] That is, individuals are subject to a series of decisions over the course of their lives, and the effects of these decisions both accumulate and compound over time. When

[10].[27] See, in particular, the quotes from David Brody.

a person receives (or is denied) one opportunity, they are likely to appear more (or less) qualified at their next encounter with a predictive system.

Machine learning systems may make self-fulfilling predictions

Suppose we find that chess skill is correlated with productivity among software engineers. Here are a few possible explanations: 1. Chess skill makes one a better software engineer. 2. There are underlying cognitive skills that make one better at both. 3. College professors hold stereotypes about chess skills and software engineering, and steered students good at chess into computer science courses. 4. People with more leisure time were both able to pursue chess as a hobby and devote time to improving their software engineering skills.

Standard supervised learning does not distinguish between these causal paths. Regardless of the correct causal explanation, once a large swath of employers start using chess skill as a hiring criterion, they contribute to the perpetuation of the observed correlation. That is because applicants who are better at chess will have better opportunities for software engineering positions in this world, and these opportunities will allow them to develop their software engineering skills.

Machine learning automates the discovery of correlations such as the above. When we deploy those correlations as decision criteria, we alter the very phenomena that we are supposedly measuring. In other words, using non-causal variables as decision criteria may give them causal powers over time. This is not limited to machine learning: sociologists have long recognized that stereotypes that are used to justify discrimination may in fact be produced by that discrimination.[30]

Algorithmic recommendation systems may contribute to segregation

Even small preferences for homogeneous neighborhoods can lead to dramatic large-scale effects. In the Appendix, we discuss a toy model of residential segregation showing such effects. But what about the online world, e.g., online social networks? The phenomenon of people making friends with similar others (online or offline) is called homophily.

In the early days of social media, there was a hope—now seen as naive—that in the online sphere there would be no segregation due to the ease with which people can connect with each other. Instead, we observe similar patterns of homophily and segregation online as offline. This is partly because real-world relationships are reflected online, but in part it is because segregation emerges through our online preferences and behaviors.[31]

As social media has matured, concerns arising from homophily have expanded from demographic segregation to ideological echo chambers. The causal mechanisms behind polarized online discourse and the role of recommendation algorithms are being researched and debated (see the Testing chapter), but there is no doubt that online media can have structural effects.

Machine learning may lead to homogeneity of decision making

If a company hires only people whose names begin with certain letters of the alphabet, it may seem absurd but not necessarily a cause for alarm. One reason behind this intuition is that we expect that the effect of any such idiosyncratic policies will cancel out, given that job candidates have many firms to apply to. If, on the other hand, every employer adopted such a policy, then the experience of job seekers becomes radically different.

Machine learning results in more homogeneous decision making compared to the vagaries of individual decisions. Studies of human behavior show that human decisions have a lot of "noise".[11] Removing the noise is one of the main attractions of statistical decision making. But there are also risks. If statistical decision making results in similar decisions being made by many decision makers, otherwise idiosyncratic biases could become amplified and reified to the point where they create structural impediments.[33]

Homogeneity can happen in many ways. At a high level, if many machine learning systems use the same training data and the same target variable, they will make roughly the same classifications, even if the learning algorithms are very different. Intuitively, if this *weren't* the case, one could make more accurate classifications by ensembling their predictions. For a stark illustration of homogeneous predictions from the domain of predicting life outcomes, see the Fragile Families Challenge.[34]

Alternatively, many decision makers could use the same underlying system. There are anecdotes of job seekers being repeatedly screened out of jobs on the basis of personality tests, all offered by the same vendor.[35]

Even *individual* algorithmic systems may have such an outsized influence in society that their policies may have structural effects. The most obvious example are systems adopted by the state, such as a predictive policing system that leads to the overpolicing of low-income neighborhoods.

But it is private platforms, especially those with a global scale, where this effect has been most prominent. Take content moderation: a small number of social media companies together determine which types of speech can be a part of mainstream online discourse and which communities are able to mobilize online. Platform companies have faced criticism for allowing content that incites violence and, conversely, for being overzealous in deplatforming individuals or groups.

In some cases, platform policies are shaped by the capabilities and limitations of machine learning.[36] For example, algorithms are relatively good at detecting nudity but relatively poor at detecting context. Companies such as Facebook have had broad bans on nudity without much attention to context, often taking down artwork and iconic historical images.

[11] See.[32] The article makes both a descriptive claim about the inconsistency of human decisions as well as a normative claim that inconsistent decision making is poor decision making. The latter claim can be contested along many lines, one of which we pursue here.

Machine learning shifts power

Like all technologies, machine learning shifts power. To make this more precise, we analyze the adoption of machine learning by a bureaucracy. We don't mean the term bureaucracy in its colloquial, pejorative sense of an inefficient, rule-bound government agency. We rather use the term as social scientists do: a bureaucracy is a public or private entity in which highly-trained workers called bureaucrats, operating in a hierarchical structure, make decisions in a way that is constrained by rules and policies but also requires expert judgment. Firms, universities, hospitals, police forces, and public assistance programs are all bureaucracies to various degrees. Most of the decision making scenarios that motivate this book are situated in bureaucracies.

To understand the effect of adopting machine learning, we consider five types of stakeholders: decision subjects, the people who provide the training data, domain experts, machine learning experts, and policy makers. Our analysis builds on a talk by Kalluri.[37]

Machine learning as generally implemented today shifts power away from the first three categories. By representing decision subjects as standardized feature vectors, statistical decision making removes their agency and ability to advocate for themselves. In many domains, notably the justice system, this ability is central to the rights of decision subjects. Even in a relatively less consequential domain such as college admissions, the personal statement provides this ability and is a key component of the evaluation.

People who provide training data may have *knowledge* about the task at hand, but provide only their *behavior* as input to the system (think of email recipients clicking the "spam" button). Machine learning instead constructs a form of knowledge in a centralized way. In contrast, domain experts learn in part from the knowledge and lived experience of the individuals they interact with. Admittedly, experts such as physicians are often criticized for devaluing the knowledge and experience of decision subjects (patients). But the fact that such a debate is happening at all is evidence of the fact that patients have at least some power in the traditional system.

The role of domain experts is also more limited compared to traditional decision making where the discretion and judgment of such experts holds sway. In machine learning applications, domain experts have two main roles: formulating the problem and task, and labeling training examples.

These effects are not always harmful. In government bureaucracies, the power wielded by "street-level bureaucrats" such as police officers and social service caseworkers—the people who translate policy into individual decisions—can be abused, and removing their discretion is often seen as a fairness intervention. Yet the discretion and human intelligence of these decision makers can also be a vital fairness-promoting element due to the existence of extenuating factors or novel circumstances not seen in the training data or covered in existing policies.[38,39]

Machine learning experts, of course, tend to have a central role. Stakeholders' requirements have to be translated into implementation by these experts; whether intentionally or unintentionally, there are often substantial gaps between the desired

policy and the policy that's realized in practice.[40] In every automated system, there is something lost in the translation of policy from human language to computer code. For example, there have been cases where software miscalculated prison inmates' eligibility for early release, with harrowing consequences including being held in prison too long and being returned to prison after being released? **(author?)** [41]]jenkins2021whistleblowers. But in those classic automated systems, these gaps tend to be mistakes that are generally obvious upon manual inspection (not that it is any comfort to those who are harmed). But when machine learning is involved, the involvement of the expert is often necessary even to recognize that something has gone wrong. This is because the policy tends to be more ambiguous (what does "high risk" mean?) and because deviations from the policy become apparent only in aggregate.

Finally, machine learning empowers policy makers or centralized decision makers. Consider a risk prediction tool used by a child protection agency to screen calls. Depending on the agency's budget and other factors, the decision maker may want to screen in a higher or lower proportion of calls. With a statistical tool, such a policy change can be implemented instantly, and is enormously easier than the alternative of retraining hundreds of case workers to adjust their mental heuristics. This is just one example that illustrates why such tools have proven so attractive to those who make the decision to deploy them.

Structural interventions for fair machine learning

The fact that machine learning may contribute to structural discrimination motivates the need for interventions that are similarly broad in scope. We call these structural interventions: changing the way machine learning gets built and deployed. The changes we have in mind go beyond the purview of any single organization, and require collective action. This could take the form of a broad social movement, or other collectives including communities, workers, researchers, and users.

Reforming the underlying institutions

One approach is to focus on the underlying institution rather than the technology, and change it so that it is less prone to adopt harmful machine learning tools in the first place. For example, shifting the focus of the criminal justice system from incapacitation to rehabilitation could decrease the demand for risk prediction tools.[42] Many scholars and activists distinguish between reform and abolition (sometimes called non-reformist reform), abolition being a more radical and transformative approach [12]. For our purposes, however, they both have the effect of centering the intervention on the institution rather than the technology.

[12].[43] See also Chapter 5 of[44] which includes a discussion of tech tools for resisting oppressive institutions.

In many domains, the very purposes and aims of our institutions remain contested. For example, what are the goals of policing? Commonly accepted goals include deterrence and prevention of crime, ensuring public safety and minimizing disorder, and bringing offenders to justice; they might also include broader efforts to improve the health and vitality of communities. The relative importance of these goals varies between communities and over time. Thus, formulating police allocation decisions as an optimization problem, as predictive policing systems do, involves taking positions on these deeply contested issues.

History shows us that many institutions that may feel like fixtures of modern society, such as higher education, have in fact repeatedly redefined their goals and purposes to adapt to a changing world. In fact, sometimes the impetus for such shifts was to *more effectively discriminate*. In the early twentieth century, elite American universities morphed from treating size (in terms of enrollment) as a source of prestige to selectivity. A major reason for this change was to curtail the rising proportion of Jewish students without having to introduce explicit quotas; the newfound mission of being selective enabled them to emphasize traits like character and personality in admissions, which in turn allowed much leeway for discretion. In fact, this system that Harvard adopted in 1926 was the origin of the holistic approach to admissions that continues to be contentious today.[45]

Some scholars have gone beyond the position that intervention to address algorithmic harms should focus on the underlying institution, and argued that the adoption of automated decision making actually enables resistant institutions to stave off necessary reform. Virginia Eubanks examines four public assistance programs for poor people in the United States—food assistance, Medicaid, homelessness, and at-risk children.[46] In each case there are eligibility criteria administered automatically, some of which use statistical techniques. The book documents the harmful effects of these systems, including the punitive effects on those deemed ineligible; the disproportionate impact of those burdens on low-income people of color, especially women; the lack of transparency and seeming arbitrariness of the decisions; and the tracking and surveillance of the lives of poor that is necessary for these systems to operate.

These problems may be fixable to some extent, but Eubanks has a deeper critique: that these systems distract from the more fundamental goal of eradicating poverty ("We manage the individual poor in order to escape our shared responsibility for eradicating poverty"). In theory, the two approaches may coexist. In practice, Eubanks argues, these systems legitimize the idea that there is something wrong with some people, hide the underlying structural problem, and foster inaction. They also incur a high monetary cost that could otherwise be put toward more fundamental reform.

Community rights

Harmful technologies are often legally justified under a notice-and-consent framework which rests on an individualistic conception of rights and is ill-equipped to address collective harms. For example, police departments obtain footage en

masse from residential security cameras with the consent of residents through centralized platforms like Amazon Ring.[47] However, consent is not a meaningful check in this scenario, because the people who stand to be harmed by police abuse of surveillance footage—such as protesters or members of racial minorities who had the police called on them for "acting suspiciously"—are not the ones whose consent is sought or obtained.

This gap is especially salient in machine learning applications: even if a classifier is trained on data provided with consent, it may be applied to nonconsenting decision subjects. An alternative is to allow groups, such as geographic communities, the right to collectively consent to or reject the adoption of technology tools. In response to the police use of facial recognition, civil liberties activists advocated for a community right to reject such tools; the success of this advocacy has led to various local bans and moratoria.[48] In contrast, consider online targeted advertising, another technology that has faced widespread dissent. In this case, there are no analogous collectives who can organize effective resistance, and hence attempts to reject the technology have been much less successful.[49]

Beyond collective consent, another goal of community action is to obtain a seat at the table in the design of machine learning systems as stakeholders and participants whose expertise and lived experience shapes the conception and implementation of the system rather than mere data providers and decision subjects. Among other benefits, this approach would make it easier to foresee and mitigate representational harms—issues such as demeaning categories in computer vision datasets or image search results that represent offensive stereotypes. But there are also potential risks to participatory design: it may create further burdens for members of underrepresented communities, and it may act as a smokescreen for organizations resisting meaningful change. It is essential that participation be recognized as labor and be fairly compensated.[50]

Regulation

Regulation that promotes fair machine learning can take the form of applying existing laws to decision systems that incorporate machine learning, or laws that specifically address the use of technology and its attendant harms. Examples of the latter include the above-mentioned bans on facial recognition, and restrictions on automated decision making under the European Union's General Data Protection Regulation (GDPR). Both flavors of regulation are evolving in response to the rapid adoption of machine learning in decision making systems. Regulation is a major opportunity for structural intervention for fair machine learning. Yet, because of the tendency of law to conceptualize discrimination in narrow terms, its practical effect on curbing harmful machine learning largely remains to be seen.[51]

The gap between the pace of adoption of machine learning and the pace of law's evolution has led to attempts at self-regulation: a 2019 study found 84 AI ethics guidelines around the world.[52] Such documents don't have the force of law but attempt to shape norms for organizations and/or individual practitioners. While self-regulation has been effective in some fields such as medicine, it is doubtful if AI

self-regulation can address the thorny problems we have identified in this chapter. Indeed, industry self-regulation generally aims to forestall actual regulation and the structural shifts it may necessitate.[13]

Workforce interventions

Machine learning shifts power to machine learning experts, which makes the ML workforce an important locus of interventions. One set of efforts is aimed at enabling more people to benefit from valuable job opportunities in the industry[54] and to fight imbalances of power within the workforce—notably, between technology experts and those who perform other roles such as annotation.[55] Another set of efforts seeks to align the uses of ML with ethical values of the ML workforce. The nascent unionization movement in technology companies seems to have both objectives.

While a more diverse workforce is morally valuable for its own sake, it is interesting to ask what effect it has on the fairness of the resulting products. One experimental study of programmers found that the gender or race of programmers did not impact whether they produced biased code.[56] However, this is a lab study and should not be seen as a guide to the effects of structural interventions. For example, one causal path by which workforce diversity could impact products (not captured in the study's design) is that a team with a diversity of perspectives may be more willing to ask critical questions about whether a product should be built or deployed.

Another workforce intervention is education and training. Ethics education for computer science students is on the rise, and a 2018 compilation included over 200 such courses.[57] A long-standing debate is about the relative merits of standalone courses and integration of ethics into existing computer science courses.[58] Professional organizations such as the Association for Computing Machinery (ACM) have had codes of ethics for several decades, but it's unclear if these codes have had a meaningful impact on practitioners.

In many professional fields including some engineering fields, ethical responsibilities are enforced in part through licensing of practitioners. Professionals such as doctors and lawyers must master a body of professional knowledge, including ethical codes, are required by law to pass standardized exams before being licensed to practice, and may have that license revoked if they commit ethical transgressions. This is not the case for software engineering. At any rate, the software engineering certification standards that do exist[59] have virtually no overlap with the topics in this book.

The research community

The machine learning research community is another important locus for reform and transformation. The most significant push for change has been the ongoing fight for treating research topics such as fairness, ethics, and justice as legitimate

[13] For a deeper critique of industry-led statements of principles see.[53]

Type	Intervention	Example
Modifying the outputs	Reallocation	Group-specific decision thresholds
Modifying the decision process	Combatting interpersonal discrimination	Implicit bias training
	Formalization	Adopting statistical decision making
	Procedural protections	Explanation and recourse
Before the decision	Outreach	Sending mailers about scholarships
	Intervening on causal factors	Job training, preventive health
After the decision	Modifying the environment	Helping defendants show up to court

Figure 2: A summary of major types of organizational interventions

and first-rate. Traditionally, a few topics in machine learning such as optimization algorithms have been considered "core" or "real" machine learning, and other topics—even dataset construction—seen as peripheral and less intellectually serious.

A few other key debates: should all machine learning researchers be required to reflect on the ethics of their research?[60] Is there too much of a focus on fixing bias as opposed to deeper questions about power and justice[61]? How to center the perspectives of people and communities affected by machine learning systems? What is the role of industry research on fair machine learning given the conflicts of interest?

Organizational interventions for fairer decision making

The structural interventions we've discussed above require social movements or other collective action and have been evolving on a timescale of years to decades. This is not to say that an organization should throw up its hands and wait for structural shifts. A plethora of interventions are available to most types of decision makers. This section is an overview of the most important ones.

As you read, observe that the majority of interventions attempt to improve outcomes for all decision subjects rather than viewing fairness as an inescapable tradeoff. One reason this is possible is that many of them don't operate at the moment of decision. Note, also, that evaluating the effects of interventions—whether with respect to fairness or other metrics—generally requires causal inference. Finally, only a small subset of potential fairness interventions can be implemented in the framework of machine learning. The others focus on organizational or human practices rather than the technical subsystem involved in decision making.

Redistribution or reallocation

Redistribution and reallocation are terms that refer to interventions that modify a decision-making process to introduce an explicit preference for one or more groups, usually groups considered to be disadvantaged. When we talk about fairness interventions, this might be the kind that most readily comes to mind.

When applied to selection problems where there is a relatively static number of slots, as is typical in hiring or college admissions, a plethora of algorithmic fairness interventions reduce to different forms of reallocation. This includes techniques such as adding a fairness constraint to the optimization step, or a post-processing adjustment to improve the scores of the members of the disadvantaged groups. This is true regardless of whether the goal is demographic parity or any other statistical criterion.

Reallocation is appealing because it doesn't require a causal understanding of why the disparity arose in the first place. By the same token, reallocation is a crude intervention. It is designed to benefit a group—and it has the advantage of providing a measure of transparency by allowing a quantification of the group benefit—but most reallocation procedures don't incorporate a notion of deservingness of members within that group. Often, reallocation is accomplished by a uniform preference for members of the disadvantaged group. Alternatively, it may be accomplished by tinkering with the optimization objective to incorporate a group preference. In this approach, distributing the fruits of reallocation within the group is delegated to the model, which may end up learning a non-intuitive and unintended allocation (for example, an intersectional subgroup may end up further disadvantaged compared to a no-intervention condition). At best, reallocation methods will aim to ensure that relative ranking within groups is left unchanged.

As crude as reallocation is, another intervention with an even worse tradeoff is to omit features correlated with group identity from consideration. To be clear, if the feature is statistically, causally, or morally irrelevant, that may be a good reason for omitting it (Chapter 2). But what if the feature is in fact relevant to the outcome? For example, suppose that people who contribute to open-source software projects tend to be better software engineers. This effect acts through a morally relevant causal path because programmers obtain useful software-engineering skills through open-source participation. Unfortunately, many open-source communities are hostile and discriminatory to women and minorities (this is perhaps because they lack the formal organizational structures that firms use to keep interpersonal discrimination in check to some degree). Recognizing this bias, a software company could either explicitly account for it in hiring decisions or simply omit consideration of open-source contributions as a criterion. If it does the latter, it ends up with less qualified hires on average; it also disadvantages the people who braved discrimination to develop their skills, arguably the most deserving group.

Omitting features based on statistical considerations without a moral or causal justification is extremely popular in practice because it is simple to implement, politically palatable, and avoids the legal risk of disparate treatment.

Combatting interpersonal discrimination

Rather than intervene directly on the outputs, organizations can try to improve the process of decision making. In many cases, discriminators are surprisingly candid about their prejudices in surveys and interviews.[62] Can they perhaps be trained out of their implicit or overt biases? This is the idea behind prejudice reduction, often called diversity training.

But does diversity training work? Paluck & Green conducted a massive review of nearly a thousand such interventions in 2009.[63] The interventions include promoting contact with members of different groups, recategorization of social identity, explicit instruction, consciousness raising, targeting emotions, targeting value consistency and self-worth, cooperative learning, entertainment (reading, media), discussion and peer influence. Unfortunately, only a small fraction of the published studies reported on field experiments; Paluck & Green are dubious about both observational field studies and lab experiments. Overall, the field experiments don't provide much support for the effectiveness of diversity interventions. That said, there were many promising lab methods that hadn't yet been tested in the field. A more recent review summarizes the research progress from 2007 to 2019.[64]

Minimizing the role of human judgment via formalization

Approaches like implicit bias training seek to improve the judgment of human decision makers, but ultimately defer to that judgment. In contrast, formalization aims to curb judgment and discretion.

The simplest formalization technique is to withhold the decision subject's identity (or other characteristics considered irrelevant) from the decision maker. Although this idea dates to antiquity, in many domains the adoption of anonymous evaluation is a recent phenomenon and has been made easier by technology.[65] Two major limitations of this approach are the ubiquitous availability of proxies and the fact that anonymization is not feasible in many contexts such as in-person hiring interviews.[14]

A more ambitious approach is rule-based or statistical decision making that removes human discretion entirely. For example, removing lender discretion in loan underwriting was associated with a nearly 30% increase in the approval rates of minority and low-income applicants, while at the same time increasing predictive accuracy (of the risk of default).[67] Human decision makers tend to selectively ignore credit history irregularities of White applicants.[68]

In some ways, machine learning can be seen as a natural progression of the shift from human judgment to rule-based decision making. In machine learning, the discovery of the rule—and not just its application—is deferred to the data and implemented by an automated system. Based on this, one might naively hope that machine learning will be even more effective at minimizing discrimination.

[14] Even in these contexts, blinding of attributes that are not readily inferrable can be effective. Indeed, it is frowned upon to inquire about candidates' marital status during job interviews, and such inquiries may be treated as evidence of intent to discriminate.[66]

However, there are several counterarguments. First, claims of the superiority of statistical formulas over human judgment, at least in some domains, have been questioned as being based on apples-to-oranges comparisons because the human experts did not view their role as pure prediction. For example, judges making sentencing decisions may consider the wishes of victims, and may treat youth as a morally exculpatory factor deserving of leniency.[69] Second, there has been a recognition of all the ways in which machine learning can be discriminatory, which is of course a central theme of this book. Third, there are numerous potential drawbacks such as a loss of explainability and structural effects that are not captured by the human-machine comparisons.

Perhaps most significantly, incomplete formalization can simply shift the abuse of discretion elsewhere. In Kentucky, the introduction of pretrial risk assessment *increased* racial disparities for defendants with the same predicted risk. The effect appears to be partly because of differential adoption of risk assessment in counties with different racial demographics, and partly because even the same judges are more likely to override the recommended decision for Black defendants compared to White defendants.[70,71] In Ontario, social service caseworkers described how they manipulate the inputs to the automated system to get the outcomes they want.[72][15] In Los Angeles, police officers used many strategies to resist being managed by predictive policing algorithms.[73]

The most pernicious effect of formalization as a fairness intervention is that it may shift discretion to earlier stages of the process making bias *harder* to mitigate. Examples abound. Mandatory minimum sentencing guidelines for drug possession in the United States in the 1980s were justified in part as a way to combat judges' biases and arbitrariness,[74] but are now widely recognized as overly punitive and structurally racist. One way in which such laws can encode race is the 100-to-1 sentencing disparity between powder and crack cocaine, the popularity of the two forms of the same drug differing by income and socioeconomic status.[75] A very different kind of example comes from Google, which has had a vaunted, highly formalized process for recruiting in order to combat bias and enhance the quality of decisions.[76] But recruiters have argued that this process in fact bakes in bias because it incorporates a ranking of colleges in which Historically Black Colleges and Universities are not ranked at all.[77]

The Harvard admissions lawsuit from Chapter 5 is another case study of formalization versus holistic decision making. Plaintiffs point out that the admissions criteria include subjective assessments of personality traits such as likability, integrity, helpfulness, kindness, and courage. Harvard scored Asian-American applicants on average far lower on these traits than any other racial group. Harvard, on the other hand, argues that evaluating the "whole person" is important to identify those with unique life experiences that would contribute to campus diversity, and that a consideration of subjective traits is a necessary component of

[15] Caseworkers report doing so in order to work around the limitations and non-transparency of the automated system to achieve just outcomes for clients. The difficulty of distinguishing between abuse of discretion and working around an overly rigid system further illustrates the double-edged nature of formalization as a fairness intervention.

this evaluation.

Procedural protections

Diversity training and formalization are examples of procedural fairness interventions. There are many other procedural protections: notably, making the process transparent, providing explanations of decisions, and allowing decision subjects to contest decisions that may have been made in error. As we discussed above, procedural protections are more important when machine learning is involved than for other types of automated systems.

United States law emphasizes procedural fairness over outcomes. This is one reason for the great popularity of diversity training despite its questionable effectiveness.[78] When the decision maker is the government, the legal conception of fairness is even more focused on procedure. For example, there is no notion of disparate impact under United States constitutional law.

While some procedural interventions such as diversity training have been widely adopted, many others remain rare despite their obvious fairness benefits. For example, few employers offer candid explanations for job rejection. Decision makers turning to automated systems are often looking to cut costs, and may hence be especially loath to adopt procedural protections.

There are many examples of fairness concerns with automated systems for which *only* procedural protections can be an effective remedy (other than scrapping the system altogether). For example, Google's policy is to suspend users across its entire suite of services if they violate its terms of service. There are many anecdotal reports from users who have lost years' worth of personal and professional data, insist that Google's decision was made in error, and that Google's appeal process did not result in a meaningful human review of the decision.

Outreach

The rest of the interventions are not about changing the decision making process (or outcomes). Instead, they change something about the decision subjects or the organizational environment.

A recent study sought to address the puzzling phenomenon that low-income students tend not to attend highly selective colleges, even when their strong academic credentials qualify them for admission and despite the availability of financial aid that would make it *cheaper* to attend a selective institution.[79] The authors designed an intervention in which they sent flyers to low-income high-school students informing them about a new scholarship at the University of Michigan, and found that compared to a control group, these students were more than twice as likely to apply as well as enroll at the University. The effect was entirely due to students who would have otherwise attended less selective colleges or not attended college at all. The targets of outreach were highly qualified students identified based on standardized test scores (ACT and SAT), which allowed the university to guarantee financial aid conditional on admission. It is

worth reiterating that this was a purely informational intervention: the scholarship was equally available to students in the control group, who received only postcards listing University of Michigan application deadlines.

To the extent that disparities are due to disadvantaged groups lacking knowledge of opportunities, informational interventions should decrease those disparities, but this point doesn't appear to be well-researched. For example, the Michigan study targeted the intervention at low-income students, so it doesn't address the question of whether informing *all* students would close the income gap.

Intervening on causal factors

If we understand the causal factors that lead to underperformance of some individuals or groups, we can intervene to mitigate them. Like informational interventions, this approach seeks to help all individuals rather than simply minimize disparities. This type of intervention is extremely common. Some examples: job training programs for formerly incarcerated people to improve welfare and decrease the chances of recidivism; efforts to bolster math and science education to address an alleged labor shortage of engineers (a so-called pipeline problem); and essentially all of public health and preventive healthcare. The use of randomized controlled trials to identify and intervene on the causes of poverty has been so influential in development economics that it led to the 2019 Nobel Prize to Duflo, Banerjee, and Kremer.

In a competitive market, such as an employer competing for workers, this intervention may not pay off for an individual decision maker from an economic perspective: job seekers who have benefited from the intervention may choose to join other firms instead. Many approaches have been used to overcome this misalignment of incentives. Firms may act collectively, or the state may fund the intervention. If a firm is large enough, the overall payoffs could be so high relative to the cost of the intervention that the reputational benefit to the firm may be sufficient to justify it.

Modifying the organizational environment

If decision makers have many opportunities to intervene before the point of decision (e.g. hiring), they also have opportunities to intervene after that point to ensure that individuals fulfill their potential. If a firm finds that few minority employees are successful, it may be because the workplace is hostile and discriminatory.

In other cases, some individuals or groups may need additional accommodations to remedy past disadvantages or because of morally irrelevant differences. A few examples: remedial courses for disadvantaged students, a peer group for first-time college students, need-based scholarships, a nursing mother's room in a workplace, and disability accommodations.

Accommodation isn't simply redistribution in disguise: it does not (or need not) involve an explicit preference for the disadvantaged group. Even if the accommodation is made available to everyone, the disadvantaged group will preferentially

benefit from it. This is obvious in the case of, say, disability accommodations. In other cases this is less obvious, but no less true. Even if financial aid were available to all students at a university, it would differentially benefit low-income students.

However, the actual effects of accommodations can be hard to predict and must be carefully measured empirically. A notable example comes from a study showing that men benefit from gender-neutral clock-stopping policies.[80] Such policies in universities allow both men and women to add time to the tenure clock with the birth of a child. While they are often adopted in the interest of fairness, the study shows that they increase men's tenure rates and lower women's; this is presumably because men are able to be more productive during their extended time due to differences in child-care responsibilities or the impact of the birth itself.[16]

Here's a stark example of how organizational policies can cause people to fail and how easily they can be remedied. In New York City, there are approximately 300,000 cases of low level offenses every year. The defendants are required to appear in court;[17] if they fail to appear, arrest warrants are automatically issued. Historically, a remarkable 40% of defendants fail to appear in court. The resulting negative consequences of Failure to Appear (FTA) are both severe and unequally distributed: for instance, members of groups that are subject to overpolicing are more likely to be arrested. Remarkably, a study found that FTA rates decreased from 41% to 26% simply by redesigning the summons form to be less confusing and sending defendants text messages shortly before their court dates[81]!

Summary

We looked at seven broad types of fairness interventions that organizations can deploy. The majority of these interventions potentially improve opportunity for all decision subjects as they are motivated by some underlying injustice rather than merely mitigating some disparity. In fact, interventions that aim to address an underlying justice might sometimes increase certain disparities between groups—a possibility that would be morally justified under a non-comparative notion of fairness that calls for treating each subject as they ought to be treated.[82]

Comparative notions of fairness are appealing to focus on because they are easy to quantify, but we shouldn't forget the deeper questions. A domain where this seems to have happened is algorithmic hiring. Tools used in algorithmic hiring utilize situational judgment tests, personality tests, and sometimes much more dubious techniques—increasingly involving machine learning—for screening and selecting candidates. Firms adopt such tools to cut recruitment costs, especially for low-wage positions where the cost of hiring a worker through the traditional process can be seen as significant in relation to a worker's contribution to the firm's revenue over the course of the period of their employment.

These tools are problematic for many reasons. While they aim to formalize the

[16]However, note that the policy has two fairness goals: to mitigate the adverse career impact of childbirth and to decrease gender disparities in said impacts. Presumably the policy still meets the first goal even if it fails the second.

[17]For offenses of the lowest severity, the summons may be resolved by mail.

hiring process, they often use attributes that are morally and causally irrelevant to job performance. HireVue, for example, previously relied on facial expressions and intonations in a person's voice as part of its automated assessment. They also fail to take a broad view of discrimination. Focusing narrowly on minimizing disparities in hiring rates across groups leaves unaddressed what kind of environment employees will encounter once hired. If job applicants from certain groups were previously predicted to perform poorly in a certain workplace, the employer should strive to understand the reasons for this difference in success, rather than simply trying to find members of these groups that might be able to succeed under such unfavorable, unwelcoming, or hostile conditions. Parity-promiting interventions change the selection process, but preserve the organizational status quo, endorsing the idea that the candidates that have been selected should be able to deal with these conditions sufficiently well to be as productive as their peers who don't face similar challenges. Other productive—and potentially less harmful—forms of intervention include on-the-job training (which might be understood as a way of intervening on causal factors), meaningful feedback for rejected applicants (which would provide some degree of procedural protection, but also help guide applicants' future investment in their own development), and a strategic approach to sourcing candidates who firms with more accurate tools might now be better able to assess.

The narrow focus on disparities can mean that there is little consideration of the quality of decisions made by the tools. Tools that simply lack validity raise a host of normative concerns. Notably, assessments that achieve approximate demographic parity but continue to suffer from accuracy disparity (also called differential validity) can set members of certain groups up for failure by expecting them to be able to perform better than they would be currently prepared to.[18]

To reiterate, we do not advocate for treating statistical fairness criteria as constraints, at least in the first instance. That approach assumes that reallocation is the only available intervention. Instead, if we treat statistical fairness criteria as diagnostics, we are likely to uncover deeper problems that require remedying. Unfortunately, these deeper remedies are also harder. They require both causal inference and normative depth. That is of course why they are often ignored, and foundational questions remain unaddressed.

A case in point: a 2021 paper analyzes the fairness of pre-trial detention in a non-comparative sense.[83] How risky does a defendant have to be so that the expected benefit to public safety justifies the harm to the defendant from detention? Using the clever approach of asking survey recipients to choose between being detained and becoming victims of certain crimes, the authors conclude that pretrial detention is essentially never justified.

The study's method is sure to be debated, but the point remains that there have been relatively few principled, quantitative attempts to justify the risk thresholds used in pretrial detention. There have been many other calls to end pretrial detention based on different moral and legal arguments. When such foundational

[18] See also the discussion of the limitations of independence as a fairness criterion in Chapter 3.

questions continue to be debated, it would be exceedingly premature to declare a risk-based pretrial detention system to be "fair" because it satisfied some statistical criterion.

Appendix: a deeper look at structural factors

Let us briefly discuss two phenomena that help explain the long-run persistence of inequality: segregation and feedback loops.

The role of segregation

A structural factor that exacerbates all of the mechanisms of discrimination we discussed is the segregation of society along the lines of group identity. Segregation arguably enables interpersonal discrimination because increased contact among groups decreases prejudice toward outgroups—the controversial contact hypothesis.[84]

At a structural level, segregation sustains inequality because an individual's opportunities for economically productive activities depend on her social capital, including the home, community, and educational environment. A strand of the economics literature has built mathematical models and simulations to understand how group inequalities—especially racial inequalities—arise and persist indefinitely even in the absence of interpersonal discrimination, and despite no intrinsic differences between groups. In the extreme case, if we imagine two or more groups belonging to non-interacting economies that grow at the same rate, it is intuitively clear that differences can persist indefinitely. If segregation is imperfect, do gaps eventually close? This is sensitive to the assumptions in the model. In Lundberg and Startz's model the gaps close eventually, although extremely slowly.[85] In Bowles et al.'s model, they don't under some conditions;[86] one reason is that the disadvantaged group might face higher costs of labor-market skill acquisition due to lower social capital.[19]

In the United States, after the civil rights legislation of the 1960s and 70s, residential segregation by race has been decreasing, albeit slowly. On the other hand, residential segregation by income appears to be increasing.[88]

The role of feedback loops

There is a classic economic model of feedback loops in the context of a labor market.[89] There are two groups of workers and two types of jobs: high and low skilled, with high-skilled jobs requiring certain qualifications to perform effectively. Under suitable assumptions (especially, employers cannot perfectly observe worker qualifications before hiring them, but only after providing costly on-the-job skills training) there exists an economic equilibrium in which the following feedback loop sustains itself:

[19] A pioneering work in this area is.[87]

1. The employer practices wage discrimination between the two groups.
2. As a result, the disadvantaged group achieves lower returns to investment in qualifications.
3. Workers, assumed to be rational, respond to such a differential by investing differently in acquiring qualifications, with one group acquiring more qualifications.
4. The employer—again, under certain rationality assumptions—wage discriminates because of the observed difference in qualifications.

The significance of this model is that it can explain the persistence of inequality (and discrimination) without assuming intrinsic differences between the groups, and without employers discriminating between equally qualified workers. It should be viewed as showing only the possibility of such feedback loops. Like any theoretical model, a claim that such a feedback loop explains some actually observed disparity would require careful empirical validation.

Bibliography

[1] Cody Cook, Rebecca Diamond, Jonathan Hall, John A List, and Paul Oyer. The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. Technical report, National Bureau of Economic Research, 2018.

[2] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[3] Langdon Winner. *Do artifacts have politics?* Routledge, 2017.

[4] Robert Doyle Bullard, Glenn Steve Johnson, and Angel O Torres. *Highway robbery: Transportation racism & new routes to equity*. South End Press, 2004.

[5] Alex Rosenblat, Karen EC Levy, Solon Barocas, and Tim Hwang. Discriminating tastes: Customer ratings as vehicles for bias. *Data & Society*, pages 1–21, 2016.

[6] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(3):1–40, 2017.

[7] Mario L Small and Devah Pager. Sociological perspectives on racial discrimination. *Journal of Economic Perspectives*, 34(2):49–67, 2020.

[8] Devah Pager and Hana Shepherd. The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annu. Rev. Sociol*, 34:181–209, 2008.

[9] Andrew Altman. Discrimination. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.

[10] Lily Hu. Direct effects. https://phenomenalworld.org/analysis/direct-effects, 2020.

[11] Wikipedia contributors. Same-sex marriage — Wikipedia, the free encyclopedia, 2021.

[12] Richard Rothstein. *The color of law: A forgotten history of how our government segregated America*. Liveright Publishing, 2017.

[13] Jamie Fellner. Race, drugs, and law enforcement in the united states. *Stan. L. & Pol'y Rev.*, 20:257, 2009.

[14] Ted Mouw. Are black workers missing the connection? the effect of spatial distance and employee referrals on interfirm racial segregation. *Demography*, 39(3):507–528, 2002.

[15] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

[16] Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323):389–391, 2017.

[17] Candace West and Don H Zimmerman. Doing gender. *Gender & society*, 1(2):125–151, 1987.

[18] Felipe Valencia Caicedo. The mission: Human capital transmission, economic persistence, and culture in south america. *The Quarterly Journal of Economics*, 134(1):507–556, 2019.

[19] Melissa Dell. The persistent effects of peru's mining mita. *Econometrica*, 78(6):1863–1903, 2010.

[20] Guglielmo Barone and Sauro Mocetti. Intergenerational mobility in the very long run: Florence 1427-2011. *Bank of Italy Temi di Discussione (Working Paper) No*, 1060, 2016.

[21] Shai Davidai and Thomas Gilovich. Building a more mobile america—one income quintile at a time. *Perspectives on Psychological Science*, 10(1):60–71, 2015.

[22] Raj Chetty, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2):711–783, 2020.

[23] Rakesh Kochhar and Anthony Cilluffo. Key findings on the rise in income inequality within america's racial and ethnic groups. *Pew Research Center*, 2018.

[24] Jessica L Semega, Kayla R Fontenot, and Melissa A Kollar. Income and poverty in the united states: 2016. *Current Population Reports*, (P60-259), 2017.

[25] Rolf Pendall and Carl Hedman. Worlds apart: Inequality between america's most and least affluent neighborhoods. *Urban Institute*, 2015.

[26] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[27] Jeremy Merill. Does Facebook still sell discriminatory ads?, 2020.

[28] Oscar H Gandy. *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Routledge, 2016.

[29] Deborah Hellman. Sex, causation, and algorithms: How equal protection prohibits compounding prior injustice. *Washington University Law Review*, 98(2):481–523, 2020.

[30] Gunnar Myrdal. *An American Dilemma: The Negro Problem and Modern Democracy, Volume 2*. Routledge, 2017.

[31] D Boyd. White flight in networked publics: How race and class shaped american teen engagement with myspace and facebook. nakamura l, chow-white pa, eds. race after the internet. *Race after the Internet*, pages 203–222, 2012.

[32] Daniel Kahneman, AM Rosenfield, L Gandhi, and T Blaser. Noise: How to overcome the high, hidden cost of inconsistent decision makinghttps. *Harvard Business Review*, 2016.

[33] Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems. *Virginia Public Law and Legal Theory Research Paper*, (2021-13), 2021.

[34] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.

[35] Cathy O'Neil. How algorithms rule our working lives. *The Guardian*, 16, 2016.

[36] Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020.

[37] Ria Kalluri. The values of machine learning. NeurIPS Queer in AI workshop, 2019.

[38] Michael Lipsky. *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation, 2010.

[39] Ali Alkhatib and Michael Bernstein. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[40] Danielle Keats Citron. Technological due process. *Wash. UL Rev.*, 85:1249, 2007.

[41] US prisoners released early by software bug. BBC news, 2015.

[42] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76. PMLR, 2018.

[43] Amna Akbar. An abolitionist horizon for police (reform). *California Law Review*, 108(6), 2020.

[44] Ruha Benjamin. *Race after Technology*. Polity, 2019.

[45] Jerome Karabel. *The chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton*. Houghton Mifflin Harcourt, 2005.

[46] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[47] Drew Harwell. Doorbell-camera firm ring has partnered with 400 police forces, extending surveillance concerns. *Washington Post*, 2019.

[48] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. *AI now report 2018*. AI Now Institute at New York University New York, 2018.

[49] Solon Barocas and Karen Levy. Privacy dependencies. *Wash. L. Rev.*, 95:555, 2020.

[50] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423*, 2020.

[51] Anna Lauren Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915, 2019.

[52] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[53] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii international conference on system sciences*, 2019.

[54] Sarah Judd. Activities for building understanding: How ai4all teaches ai to diverse high school students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 633–634, 2020.

[55] Announcing the contract worker disparity project. Tech Equity Collaborative, 2021.

[56] Bo Cowgill, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. Biased programmers? or biased data? a field experiment in operationalizing ai ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 679–681, 2020.

[57] Casey Fiesler, Natalie Garrett, and Nathan Beard. What do we teach when we teach tech ethics? a syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 289–295, 2020.

[58] C Dianne Martin, Chuck Huff, Donald Gotterbarn, and Keith Miller. Implementing a tenth strand in the cs curriculum. *Communications of the ACM*, 39(12):75–84, 1996.

[59] Wikipedia contributors. Certified software development professional — Wikipedia, the free encyclopedia, 2021.

[60] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. Unpacking the expressed consequences of ai research in broader impact statements. *arXiv preprint arXiv:2105.04760*, 2021.

[61] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proc. 4th Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 576–586, 2021.

[62] Kathryn M Neckerman and Joleen Kirschenman. Hiring strategies, racial bias, and inner-city workers. *Social problems*, 38(4):433–447, 1991.

[63] Elizabeth Levy Paluck and Donald P Green. Prejudice reduction: What works? a review and assessment of research and practice. *Annual review of psychology*, 60:339–367, 2009.

[64] Elizabeth Levy Paluck, Roni Porat, Chelsey S Clark, and Donald P Green. Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, 72, 2020.

[65] Alex Chohlas-Wood, Joe Nudell, Zhiyuan Jerry Lin, Julian Nyarko, and Sharad Goel. Blind justice: Algorithmically masking race in charging decisions. Technical report, Technical report, 2020.

[66] Pre-employment inquiries and marital status or number of children. U.S. Equal Employment Opportunity Commission, 2021.

[67] Susan Wharton Gates, Vanessa Gail Perry, and Peter M Zorn. Automated underwriting in mortgage lending: Good news for the underserved? *Housing Policy Debate*, 13(2):369–391, 2002.

[68] Gregory D Squires et al. *Capital and communities in black and white: the intersections of race, class, and uneven development*. Suny Press, 1994.

[69] Megan T Stevenson and Jennifer L Doleac. Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440*, 2019.

[70] Megan Stevenson. Assessing risk assessment in action. *Minn. L. Rev.*, 103:303, 2018.

[71] Alex Albright. If you give a judge a risk score: evidence from kentucky bail decisions. *Harvard John M. Olin Fellow's Discussion Paper*, 85, 2019.

[72] Jennifer Raso. Displacement as regulation: New regulatory technologies and front-line decision-making in ontario works. *Canadian Journal of Law and Society*, 32(1):75–95, 2017.

[73] Sarah Brayne. *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press, USA, 2020.

[74] Marvin E Frankel. Criminal sentences: Law without order. 1973.

[75] Joseph J Palamar, Shelby Davies, Danielle C Ompad, Charles M Cleland, and Michael Weitzman. Powder cocaine and crack use in the united states: An examination of risk for arrest and socioeconomic disparities in use. *Drug and alcohol dependence*, 149:108–116, 2015.

[76] Google re:work team. Guide: Hire by committee. https://rework.withgoogle.com/print/guides/6053596147744768/, 2021.

[77] Natasha Tiku. Google's approach to historically black schools helps explain why there are few black engineers in big tech. Washington Post, 2021.

[78] Lauren B Edelman. Law at work: The endogenous construction of civil rights. In *Handbook of employment discrimination research*, pages 337–352. Springer, 2005.

[79] Susan Dynarski, CJ Libassi, Katherine Michelmore, and Stephanie Owen. Closing the gap: The effect of a targeted, tuition-free promise on college choices of high-achieving, low-income students. Technical report, National Bureau of Economic Research, 2018.

[80] Heather Antecol, Kelly Bedard, and Jenna Stearns. Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? *American Economic Review*, 108(9):2420–41, 2018.

[81] Alissa Fishbane, Aurelie Ouss, and Anuj K Shah. Behavioral nudges reduce failure to appear for court. *Science*, 370(6517), 2020.

[82] Deborah Hellman. Two concepts of discrimination. *Va. L. Rev.*, 102:895, 2016.

[83] Megan T Stevenson and Sandra G Mayson. Pretrial detention and the value of liberty. *Virginia Public Law and Legal Theory Research Paper*, (2021-14), 2021.

[84] Elizabeth Levy Paluck, Seth A Green, and Donald P Green. The contact hypothesis re-evaluated. *Behavioural Public Policy*, 3(2):129–158, 2019.

[85] Shelly Lundberg and Richard Startz. On the persistence of racial inequality. *Journal of Labor Economics*, 16(2):292–323, 1998.

[86] Samuel Bowles and Rajiv Sethi. Social segregation and the dynamics of group inequality. 2006.

[87] Glenn C Loury. A dynamic theory of racial income differences. Technical report, Discussion paper, 1976.

[88] Douglas S Massey, Jonathan Rothwell, and Thurston Domina. The changing bases of segregation in the united states. *The Annals of the American Academy of Political and Social Science*, 626(1):74–90, 2009.

[89] Kenneth Arrow et al. The theory of discrimination. *Discrimination in labor markets, Achenfelter, A. Ress (eds.)*, 1973.