4

Relative notions of fairness

In Chapter 3, we considered a range of statistical criteria that help to highlight group-level differences in both the treatments and outcomes that might be brought about by the use of a machine learning model. But why should we be concerned with group-level differences? And how should we decide which groups we should be concerned with? In this chapter, we'll explore the many different normative reasons we might have to object to group-level differences. This is a subtle, but important, shift in focus from Chapter 2, where we considered some of the normative reasons why *individuals* might object to decision-making schemes that distribute desirable resources or opportunities. In this chapter, we'll focus on why we might be concerned with the uneven allocation of resources and opportunities across *specific groups* and *society overall*. In particular, we'll review the normative foundations that ground claims of discrimination and calls for distributive justice. We'll then try to connect these arguments more directly to the statistical criteria developed in Chapter 3, with the aim of giving those criteria greater normative substance.

A point of terminology: We will use the terms unfairness and discrimination roughly synonymously. Linguistically, the term discrimination puts more emphasis on the agency of the decision maker. We also specifically avoid using the terminology "disparate treatment" and "disparate impact" in this chapter as these are legal terms of art with more precise meanings and legal significance; we'll address these in Chapter 6.

Systematic relative disadvantage

Discussions of discrimination in the context of machine learning can seem odd if you consider that the very point of many machine learning applications is to figure out how to treat different people differently — that is, to discriminate between them. However, what we call discrimination in this chapter is not different treatment in and of itself, but rather treatment that systematically imposes a disadvantage on one social group relative to others. The systematicity in the differences in treatment and outcomes is what gives discrimination its normative force as a concept.

To better appreciate this point, consider three levels at which people might be subject to unfair treatment. First, a person might be subject to the prejudice of an individual decision maker — for example, a specific hiring manager whose decisions are influenced by racial animus. Second, a person might encounter systematic barriers to entering certain occupations, perhaps because members of the group to which they belong are not viewed as fit to be engineers, doctors,

lawyers, etc., regardless of their true capabilities or potential. For example, in some occupations women might have limited employment opportunities across the board due to their gender. Finally, certain personal characteristics might be an organizing principle for society overall such that members of certain groups are systematically excluded from opportunities across multiple spheres of life. For example, race and gender might limit people's access not only to employment, but to education, credit, housing, etc.

In the examples from the previous paragraph, we have relied on race and gender precisely because both have served, historically, as organizing principles for many societies; they are not just the idiosyncratic bases upon which specific employers or professions have denied members of these groups important opportunities.[1]

This helps to explain why these are characteristics of particular concern and why others might not be. For example, we might not care that a particular employer or profession has systematically rejected left-handed applicants beyond the fact that we might find the decision arbitrary and thus irrational, given that handedness might not have anything to do with job performance.

But if handedness became the basis for depriving people of opportunities across the board and not just by this one decision maker or in this one domain, we might begin to view it as problematic. To the extent that handedness dictates people's standing and position in society overall, it would rise to the level of a characteristic worthy of special concern.[2]

Race and gender — among others enumerated in discrimination law and described in more detail in Chapter 6 — rise to such a level because they have served as the basis for perpetuating systematic relative disadvantage across the board. In extreme cases, certain characteristics can provide the foundations for a strict social hierarchy in which members of different groups are slotted into more or less desirable positions in society. Such conditions can create the equivalent of a caste system,[3] in which certain groups are confined to a permanent position of relative disadvantage.[1]

It is also important to note the unique threat posed by differential treatment on the basis of characteristics that persist intergenerationally. For example, children are often assumed to belong to the same racial group as their biological parents, making the relative disadvantage that people may experience due to their race especially systematic: children born into families that have been unfairly deprived of resources and opportunities will have less access to these resources and opportunities, thereby limiting from the start of their lives how effectively they might realize their potential — even before they might be subject to discrimination themselves.

[1] Of course, such stratification need not originate from formal policies or from one or even a small handful of highly consequential decisions. Many seemingly small actions can cumulatively reinforce relative advantages and disadvantages, ranging from selectively advertising a job through word of mouth to sexist comments in the workplace (more on this in Chapter 8).

Six accounts of the wrongfulness of discrimination

Scholars have developed many normative theories to account for the wrongfulness of discrimination — specifically the wrongfulness of treating people differently according to characteristics like race, gender, or disability. While each of these theories is concerned with how such differential treatment gives rise to systematic relative disadvantage, they differ in how they understand what makes decision making on the basis of these characteristics morally objectionable.

Relevance: One reason — and perhaps the most common reason — to object to discrimination is because it relies on characteristics that bear little to no relevance to the outcome or quality that decision makers might be trying to predict or assess.[4,3] For example, one reason why it might be wrong to base employment decisions on characteristics like race or gender is that these characteristics bear no relevance to determinations about job applicants' capacity to perform on the job. Note that this is a variant of the objection that we covered in Chapter 2, where individuals might contest decisions based on the fact that they were rendered on the basis of irrelevant information. In this case, it is important not only that the reliance on irrelevant factors leads to mistakes, but that those mistakes lead to systematic relative disadvantage for particular social groups.

Generalizations: Or we might argue that the harm lies in the needlessly coarse groupings perpetuated by decisions made on the basis of race or gender, even if these can be shown to possess some statistical relevance to the decision at hand.[5] This harkens back to another idea in Chapter 2: that people deserved to be treated as individuals and assessed on their unique merits. As you'll recall, the intuitive idea of a perfectly individualized assessment is unattainable. Any form of judgment based on individual characteristics must draw on some generalizations and past experience. Yet we might still object to the coarseness of the generalizations, especially if there is obviously additional information that might provide a more granular — and thus more accurate — way to draw distinctions. For example, we might object if women were excluded as firefighters based on the assumption that women as a group are less likely to meet the strength requirements, as opposed to administering a fitness test to applicants of all genders.

Prejudice: Another common argument for why discrimination is wrongful is that it amounts to a form of prejudicial decision making, in which members of certain groups are presumed to be of inferior status. Rather than being merely a problem of relevance or granularity, as in the previous perspectives discussed in this section, it is a problem of beliefs, in which decision makers hold entire groups in lesser regard than others. For example, the problems with decisions guided by racial animus or misogyny is not merely that they may result in inaccurate predictions, but that decision makers hold these views in the first place.[6,7,1]

Disrespect: A related, but distinct, idea is that making decisions on the basis of such characteristics is wrongful when it demeans those who possess such characteristics.[1,8] On this account, the problem with discrimination is that it casts certain groups as categorically inferior to others and thus not worthy of equal respect. This objection differs from those based on prejudice because the

harm is not located in decision makers' belief about the inferiority of members of particular groups, but in what decision makers' actions communicate about the social status of the groups. For example, the problem with sexist hiring practices is not merely that they confine women to particular roles in society or that they are based on prejudiced beliefs, but that they suggest that women are inferior to men. Understood in this way, discrimination is harmful not merely to the specific person subject to an adverse decision but to the entire group to which the person belongs because it harms the group's social standing in the community.

Immutability: An entirely different argument for why discrimination is wrongful is because it involves treating people differently according to characteristics over which they possess no control. On this account, the reason we should care about differences in the treatment of, for example, people with or without a disability is because people with a disability may not have any control over their disability.[9, 10, 11] As explored in Chapter 2, decisions that rest on immutable characteristics deny people that possess these characteristics the agency to achieve different outcomes from the decision-making process, effectively condemning all such people to adverse outcomes.[[2] This amounts to wrongful discrimination specifically when the immutable characteristic in question is one whose use in decision making will give rise to systematic relative disadvantage.

Compounding injustice: Or perhaps the problem with discrimination stems from the fact that it may compound existing injustice.[12] In many ways, this is an extension of the previous argument about control, but with a specific focus on the effects of past injustice. In particular, it is an argument that people cannot be morally culpable for certain facts about themselves that are not the result of their own actions, especially if these facts are the result of having been subject to some past injustice. Failing to take into account the fact that members of certain groups might have been, in the past, subject to many of the types of problems described above could lead decision makers to feel perfectly justified in treating members of these groups differently. Yet the reason people might appear differently at the time of the decision might be some past injustice, including past discrimination.[13]

Ignoring this fact would mean that decision makers subject specific groups to worse decisions simply because they have already suffered some earlier harm. Note that this objection has nothing to do with concerns with accuracy; in fact, it suggests, as we first discussed in Chapter 2, that we might have a moral obligation to sometimes discount the effects of factors over which people have no control, even if that means making less accurate predictions. In Chapter 2, we considered this principle without any particular concern for distributional outcomes; here, we can adapt the principle to be one that accounts for the wrongfulness of discrimination by pointing out that suffering some past mistreatment due to someone's membership

[2] While this is especially problematic if the very same decisions could be made at least as accurately by relying on criteria over which people do possess some control, we might still object to decisions based on immutable characteristics even if there are no alternative means to make decisions at an equal level of accuracy. For example, it may be that genetic information is the most reliable basis upon which to make predictions about a person's future health. We might nevertheless object to the idea that people should be charged higher insurance premiums or denied a job (because they would impose greater healthcare costs on an employer) due to their genes.

in a particular group might be the very thing outside someone's control.

None of the above six accounts is a complete theory of the wrongfulness of discrimination. Some situations that we might view as obviously objectionable can be caught by certain theories, but missed by others. For example, objections to religious discrimination cannot be grounded on the idea that people lack control over their religious affiliations, but could be supported by reference to concerns with prejudice or disrespect.[14] Or to take another example: even if decision makers' actions are not prejudicial or demeaning, their decisions may still be based on irrelevant characteristics — a possibility that we'll consider in the next section. While there is unlikely to be a single answer to the question of why discrimination is wrong, these theories are still helpful because they highlight that we often need to consider many factors when deciding whether subjecting particular groups to systematic relative disadvantage is morally justified.

Intentionality and indirect discrimination

So far we have focused on why taking certain characteristics into account when making consequential decisions can be normatively objectionable. According to each of these accounts of the wrongfulness of discrimination, the harm originates from the choice to rely on these characteristics when making such decisions. But what about when decisions do not rely on these characteristics? Does removing these characteristics from the decision-making process ensure that it is fair?

An easy case is when the decision maker purposefully relies on proxies for these characteristics (e.g., relying on a person's name as a proxy for their gender) in order to indirectly discriminate against members of a specific group (e.g., women). The fact that such decisions do not consider the characteristics explicitly may not render them any less problematic, given that the decision maker only does so with the goal of treating members of these different groups differently. Thus, the wrongfulness of discrimination is not limited to the mere use of certain characteristics in decision making, but extends to any intentional efforts to subject members of specific groups to systematically disfavorable treatment, even if this is achieved via the use of proxies for such characteristics.[15] With that in mind, we might want to check whether any decision-making process leads to disparate outcomes for different groups as a way of potentially smoking out intentional discrimination pursued with the aid of proxies. If we discover disparities in outcomes, we might want to check whether the decision-making process was purposefully orchestrated to achieve this, even if the decision maker didn't seem to take these characteristics into account explicitly.

But what about decisions that are not purposefully designed to discriminate? What if the decisions are not motivated by prejudice? Does the mere fact that a decision-making process can lead to quite disparate outcomes for different groups mean that it is unfair? What if the decision maker can offer some reason for making decisions in this particular manner (e.g., that the employer needs people with a specific accounting credential and that such credentials happen to be held more

commonly among certain groups than others)?

We can extend some of the reasoning first introduced in the previous section to try to answer these questions. In this case, rather than asking whether the criteria under consideration are serving merely as proxies for some characteristic of concern, we could instead ask whether the choice of criteria can be justified by demonstrating that they actually serve the stated goals of the decision maker. Decision-making processes that do little to serve these goals, but nevertheless subject specific groups to systematically less favorable outcomes raise the same question about relevance that arise in cases of intentional and direct discrimination. If the chosen criteria lack relevance to the decision at hand, but result in systematic relative disadvantage for a specific group, then relying on them can easily become functionally equivalent to relying on group membership directly despite its lack of relevance to the decision at hand. In both cases, the reliance on irrelevant criteria is objectionable because it results in systematic relative disadvantage for particular groups.

Equality of opportunity

What about a process that produces disparities in outcomes but does, in fact, serve to advance the goals of the decision maker? This is a harder case to reason about. But before doing so, let's take a step back.

Equality of opportunity is an idea that many scholars see as the goal of limiting discrimination. Equality of opportunity can be understood in both narrow and broad terms. The narrow view holds that we should treat similar people similarly given their current degree of similarity. The broad view holds that we should organize society in such a way that people of similar ability and ambition can achieve similar outcomes. A position somewhere in the middle holds that we should treat seemingly dissimilar people similarly, on the belief that their current degree of dissimilarity is the result of something that we should discount (e.g., past injustice or misfortune). Let's tackle each view in turn, and see what each would imply about the question above.

The narrow view

Let's illustrate the narrow view of equality of opportunity with the notion of "individual fairness": that people who are similar with respect to a task should be treated similarly.[16] For now we take *similar* to mean closeness in features deemed relevant to the task at hand. Individual fairness is a comparative notion of fairness in that it asks whether there are any differences in the way that similar people are being treated. It is not directly concerned with the way members of different *groups* might be treated. Instead, the comparison is between all people as individuals, not between the members of specific groups. Of course, if we agree in advance that people's race, gender, and so forth are irrelevant to a task at hand, then satisfying individual fairness will also limit the degree to which people who differ according to these characteristics will receive different treatment. But this is only true to the

extent that such characteristics are deemed irrelevant; it is not an inherent part of the definition of individual fairness.

Individual fairness is related to consistency and some of the concerns with arbitrariness that we explored in Chapter 2. We often expect consistent treatment in the absence of differences that would seem to justify differential treatment, especially when the treatment determines access to important opportunities. These expectations can be so strong that a failure to meet them will provoke a visceral reaction: why did I not get the desired treatment or outcome even though I am seemingly similar along the relevant dimensions to someone that did?

What it means for people to be similar is not a given, though. In the philosophical literature, the common answer to this question is that people should be treated similarly to those who are understood to be similarly meritorious — that is, that people should be judged according to their abilities and ambitions.[17] This understanding is so common that the concept of equality of opportunity — in this narrow formulation — is often taken to be synonymous with the concept of a meritocracy. Access to desirable resources and opportunities should be dictated not by the social group to which someone happens to belong, but rather by the characteristics that are legitimately relevant to the institution seeking to advance its goals in allocating these resources and opportunities.

Much depends on what we decide to be the goals of the decision making process. It could be defined as maximizing some outcome of interest to the decision maker, such as job performance. When applicants who are predicted to perform similarly well on the job are treated similarly in the hiring process, it is often interpreted as meritocracy. Or we might say that a firm has a legitimate interest in hiring workers with the necessary training to effectively perform a job, so it might only hire those who have completed the necessary training. If making decisions on this basis leads to uneven hiring rates across groups, under the narrow view of equality of opportunity, the decision maker is blameless and under no obligation to adjust the decision-making process.

But note that employers could just as easily decide on goals that bear no obvious relationship to what we perceive as merit, such as hiring applicants who would be likely to accept a particularly low salary. With this target in place, decision makers would only have an obligation to treat people similarly who possess similar sensitivities to pay, not those who are likely to perform similarly well on the job. Would it make sense to describe this as a case in which employers ensure equality of opportunity?[1] This reveals that the narrow view of equality of opportunity does not dictate what the abiding normative principle should be that determines how we view people as similar; it only commands that similar people be treated similarly. It thus possesses little normative substance beyond consistency. (And even then, there may be practical limits to this principle. For example, an employer may still need to choose only one applicant among the many that are predicted to perform similarly well on the job — and those that did not get the job might not object that they had been treated unfairly.)

The people subject to decisions might have their own ideas about how to define similarity, ideas which might be very different from those of the decision maker.

This might be because they do not share the same goals as the decision maker, but it might also be because they believe that there are reasons completely independent of the goals of the decision-making process to view certain people as similar. Perhaps the reason we might view two different people as deserving of some opportunity is because they are equally likely to make the most of the opportunity or because they are equally needy — not only equally meritorious. In other words, we might judge job applicants as similar because they are likely to benefit similarly from the job, not only because they are likely to perform similarly well on the job. In many cases, the exact basis upon which we might view people as relevantly similar in any given context can be quite challenging for us to articulate because our conception of similarity might rely on varied normative considerations.

The broad view

A broad view of equality of opportunity sets aside questions about the fairness of any given decision-making process and instead focuses on the degree to which society overall is structured to allow people of similar ability and ambition to achieve similar success. This perspective has been most famously developed by philosopher John Rawls under the banner of "fair equality of opportunity". To simplify the argument considerably, the only defensible reason for why people might experience different outcomes over the course of their lives is if they possess different ability or ambition.[18]

Anything about the design of particular institutions in society that prevents such people from realizing their potential violates this broader understanding of equality of opportunity because it deprives equally deserving people of the same chance at success. For example, a society that fails to provide a means for similarly capable individuals born into different circumstances — one into a wealthy family and another into a poor family — to achieve similar levels of success would violate this understanding of equality of opportunity.

According to this view, the basic institutions that help to cultivate people's potential over the course of their lives must be structured to ensure that people of similar ability and ambition have similar chances of obtaining desirable positions in society — along with the many benefits that come with such positions. Thus, if education is an important mechanism by which people's potential might be fostered, a broad view of equality of opportunity would command that schools be funded such that students of equal ability and ambition — whether from wealthy or poor families — face the same prospects of long-term success. Thus, any advantage that such children might receive from their wealthy families must be offset by a corresponding intervention to ensure that such children from poor families may flourish to the same degree. If wealthier children benefit from a local tax base that can fund a high-quality public school, then society must put in place policies that make available similar amounts of funding to the public schools that educate poorer children.

Note that this is an intervention that aims to equalize the quality of the education to which wealthy and poor students will have access; it is not an intervention

into the admissions policy of any particular school. This helps to highlight the fact that the broad view of equality of opportunity is not really about fairness in decision making; it is about the design of society's basic institutions, with the goal of preventing unjust inequalities from arising in the first place. In theory, abiding by such a principle of equality of opportunity would result in a casteless society in which no one is permanently confined to a position of disadvantage despite having the potential to succeed under different circumstances.[19] Society would be structured to ensure social mobility for those who possess the relevant ability and ambition to achieve certain goals.

The middle view

Somewhere between the two poles we have just explored is a middle view that is narrowly concerned with the fairness of decision making, yet sensitive to the dynamics by which disadvantage might be perpetuated in society more broadly. This view holds that decision makers have an obligation to avoid perpetuating injustice.[13] Specifically, they must, to some degree, treat seemingly dissimilar people similarly when the causes of these dissimilarities are themselves problematic. For example, those who adopt this view might argue that universities should not simply rank order applicants by grade point averages or standardized test scores; instead, they must assess applicants with respect to the opportunities that applicants have been afforded over the course of their childhoods, recognizing that performance in school and standardized tests might differ according to past opportunity rather than according to innate ability and ambition.

To give an example, the state of Texas has a law guaranteeing admission to state-funded universities to all students who graduate in the top 10% of their high school class. This can be seen as in keeping with the middle view. If access to opportunity varies geographically, the 10% rule identifies individuals with ability and ambition without systematically disadvantaging those who had the misfortune of growing up without access to well-funded high schools.

This middle view differs from the broad view insofar as it accepts that students of equal potential will not receive equally high-quality education leading up to the moment when they finally apply to college. Yet it also differs from the narrow view insofar as it refuses to allow colleges to ignore what might account for applicants' current dissimilarity at the time that they submit their applications. Instead, the middle view suggests that there is some burden on colleges to attempt to compensate for the disadvantages that some applicants may have faced over their lifetimes such that they might appear less competitive than other applicants from more privileged backgrounds.

As a result, the middle view calls for interventions not at the level of the design of institutions, but at the level of the design of decision-making processes. It suggests that ensuring equality of opportunity requires assessing people as they *would have been* had they been afforded comparable opportunities in the past as other people of equal potential seeking the current opportunity. In certain respects, the middle view seems to be trying to realize the goals of the broad view via a

much more limited intervention: while the broad view would seem to demand that children from wealthier and poorer families have access to equally high-quality education throughout their lives, the middle view only seeks to compensate for the disadvantages experienced by poorer students relative to their wealthier peers at specific decision-making junctures that are thought to be particularly high-stakes — in this case, in college admissions. The middle view tends to focus on these junctures because they seem to be where there is an opportunity to greatly alter a child's life course and to allow them to much more effectively realize their potential.[20] Indeed, this is often why they are perceived as high-stakes.

While the interventions imagined by the middle view might seem narrower than those in the broad view because they do not require a more radical restructuring of the basic institutions of society, it's worth noting that the more discrete interventions of the middle view are designed to bring about much greater change than any one of the more continuous interventions required of the broad view. The middle view targets specific decisions that can create a sudden step change in people's life prospects, whereas the broad view aims to obviate the need for such dramatic interventions in decision making by ensuring equality throughout people's lives. In other words, the middle view will require sudden and substantial change at specific moments of decision making, while the broad view will require a significant redistribution of resources on an ongoing basis.

While the middle view clearly prohibits ignoring the reasons for differences in merit between people, it does not offer a clear prescription for how to take them into account. Taking it to its logical conclusion would result in interventions that seem extreme: it could require imagining people without the effects of centuries of oppression that they and their ancestors might have endured, suggesting, for instance, that a bank should approve a large loan to someone who does not in reality have the ability to repay it. That said, there are other areas of decision making where this view might seem more reasonable. For example, in employment, we might expect hiring managers to adopt a similar approach as admissions officers at universities, assessing people according to the opportunities they have been afforded, discounting certain differences in qualifications that might owe to factors outside their control, especially if these are qualifications that the employer could help cultivate on the job. The middle view has particular purchase in the case of insurance, where we really might want insurers to ignore the additional costs that they are likely to face in setting the price of a policy for someone with an expensive pre-existing condition outside the person's control. The extent to which we expect decision makers to bear such responsibility tends to be context-specific and contested. We will return to it shortly.

Tensions between the different views

There is an obvious conflict between the view that decision makers should treat people similarly according to how they appear at the time of decision making and the view that decision makers should treat people similarly according to how

they would have appeared had they enjoyed similar privileges and advantages as others of equal ability and ambition.[3] Thus, a person who at present seems to be more meritorious with respect to some opportunity might object if they are passed over in favor of someone who at present seems less meritorious — even if the decision maker believes the other person would be more meritorious than the person objecting if they had both enjoyed the same privileges and advantages.[4]

A similar tension arises in the way we might try to deal with discrimination. The narrow view of equality of opportunity suggests that the way to deal with discrimination is to ensure that decisions are only made on the basis of factors that are genuinely relevant to the task at hand. In other words, treating similar people similarly with respect to the task should, in most cases, rule out treating people differently according to their gender, race, etc. because these characteristics are not likely to be relevant to the task at hand. Thus, committing to the narrow view of equality of opportunity should help to keep these factors from entering the decision making process. In contrast, the middle view suggests that we might want to deal with discrimination by explicitly considering these characteristics when making decisions because it is likely that these characteristics would help to explain a good deal of the deprivation and disadvantage that people might have faced over the course of their lives. In other words, in order to understand how people who possess these characteristics might have appeared under counterfactual circumstances, decisions must take these characteristics into account. This again seems to set up a conflict because realizing a commitment to the middle view of equality of opportunity necessitates violating the requirements imposed by the narrow view of equality of opportunity.

John Roemer says that these tensions boil down to different views on "when the competition starts" for desirable positions in society: at what point in the course of our lives are we ultimately responsible for how we might compare to others?[21] Given that we have no control over the wealth of the families into which we are born or the quality of the education we might receive, we might discount any differences between people that owe to such differences. In other words, we might say that we don't think it's reasonable to adopt a narrow view of equality of opportunity when

[3]These tensions are sometimes expressed as a conflict between interventions that are narrowly concerned with the process by which decisions are made and interventions that are concerned with the outcomes produced by such decisions — between procedural and substantive notions of fairness. We avoid this perspective and language because the distinction easily blurs when you recognize that decision makers might sometimes make changes to the decision-making process with an eye towards their effect on outcomes. Most obviously, policies often try to avoid certain distributional outcomes by limiting the degree to which decisions can take certain factors into account. For example, in prohibiting employers from considering job applicants' gender or race, policies might not only be aiming to ensure that such decisions are made on the basis of relevant information, but aiming to reduce disparities in hiring rates along these lines.

[4]Note that there is no obvious tension between the narrow and broad view because the broad view would require that people of equal ability and ambition have received similar opportunities from a much earlier stage in their lives, thus already appearing similar by the time they arrive at the present moment of decision making. Of course, this would only be true in a society that engages in a significant amount of redistribution — a prospect to which proponents of the narrow view of equality of opportunity might strongly object.

assessing applicants to college because many of the relevant differences between applicants might not have emerged from a fair competition. In contrast, we might think that employers are justified in assessing job applicants, especially those for more senior roles that require many years of experience, according to the ability and ambition that they have demonstrated over the course of their careers. That is, we might agree that a narrow view of equality of opportunity is appropriate in this case because people who are well into their careers have had a fair chance to cultivate their ability and demonstrate their ambition. Tensions arise when there is disagreement over where this transition occurs in people's lives. Someone who has been passed over in favor of another person who seems less meritorious might consider it unfair because she thinks that whatever differences exist between the two of them have emerged during a period of fair competition between them. The decision maker might disagree, believing that the differences actually owe to advantages that the passed-over individual accrued during a period prior to the start of a fair competition.

Table 1: Some key differences between the three views of equality of opportunity.

	Goal	Intervention point	Who bears the cost of uplifting historically disadvantaged groups
Narrow view	Ensure that people who are similarly qualified for an opportunity have similar chances of obtaining it	Decision making	No one[5]
Middle view	Discount differences due to past injustice that accounts for current differences in qualifications	Decision making, especially critical life opportunities	Decision maker (who may pass on the cost to decision subjects)
Broad view	Ensure people of equal ability and ambition are able to realize their potential equally well	Government, on a continuous basis	Taxpayers

[5]Interventions under the narrow view include adopting more relevant decision criteria and collecting more data about decision subjects that helps make more accurate decisions. These interventions ultimately benefit the goals of the decision maker (to the extent that those goals are morally legitimate), so we don't view them as costs to the decision maker.

Merit and desert

Even if we buy into the middle and broad view of equality of opportunity, we may want some normative principles that allow us to decide just how far decision makers and the government must go in seeking to counteract inequality. More concretely, what differences between people actually justify differences in outcomes, if any? So far, we have tended to answer this question by describing those differences that cannot or should not serve as a justification for differences in outcomes. But it's also worth reflecting more deeply on the principles that seem to allow for — or perhaps even require — these differences in outcomes. In this section, we'll discuss two principles that help answer this question. The first, which already has played an important role in our discussion, is the principle of merit. The second is the principle of desert, meaning *that which is deserved.*

Merit plays an important role in all three views of equality of opportunity. In our discussion of the narrow view of equality of opportunity, merit is one way to establish how people are similar and thus who should be treated similarly. In the broad view, merit — in the form of abilities and ambitions — allows for the possibility that people of differing merit might not achieve comparable outcomes in life, but it also dictates the amount of support that must be provided to people who have the same potential as their more privileged peers but who would be unable to realize their potential as effectively in the absence of that support. Merit is crucial to the idea that there is a moral obligation to help people realize their potential — but no obligation to go any further than that. Finally, merit plays a similar role in the middle view insofar as decision makers are expected to evaluate people according to how meritorious they would have been under counterfactual circumstances. Understood in this way, all three views are perhaps more similar than they might first appear: each is calling for people of similar merit to have the same chances of success.

But what, exactly, is merit? Merit is not an objective property possessed by any given individual. Instead, merit concerns the qualities possessed by a specific person that are expected to help advance the goal of the institution who is offering the sought-after opportunity.[17] Thus, what makes a particular job applicant meritorious is how well that applicant is likely to advance the goals of the employer. While it is tempting to think that there are some universal answers to what makes any given job applicant more or less meritorious than others (e.g., how smart they are, how hardworking they are, etc.), this is not the case. Instead, merit, on this account, is purely a function of what an employer views as relevant to advancing its goals, whatever that might be. And different employers might have very different goals and very different ideas about what would do the most to help advance them. The goals of the employer might not be the goals that the job applicant would like the goals to be or what outsiders would want them to be. Others' conception of merit might differ from employers' because they simply have different ideas about the goals that employers should have in the first place.

The subjectiveness of this view of merit seems to be in conflict with earlier discussions of abilities and ambitions, which are presented as universal properties

that are not tied to the goals of any particular institution who is providing the sought-after opportunity. In suggesting that people of similar ability and ambition should have similar chances of obtaining desirable positions in society, there seems to be an implicit belief that people can be compared on their merits regardless of the opportunity in question. This reflects the fact that there are often widely-held and well-entrenched beliefs about the relevance of certain criteria in determining whether someone is deserving of a particular opportunity. An employer that assesses people according to their intelligence and industriousness is commonly understood to be assessing people according to their merit because these are the properties that can be safely assumed to help the employer advance its own interest. But there is no reason why these need to be the properties according to which job applicants must be assessed in order to ensure that the employer's decisions are based on merit.

This observation anticipates a related notion: desert. Unlike merit, desert is not tied to how well a person might help to advance the interests of the decision maker, but how well a person has performed along the dimensions that a decision maker is *expected* to evaluate people. For example, we might say that a person who has worked diligently throughout school to obtain high grades is more meritorious with respect to a job opportunity than a person who blew off classes and received middling grades, even if both people are likely to advance the goals of an employer equally well. In this account, people deserve certain opportunities given that they might have good reason to believe that certain investments would help them gain access to the sought-after opportunity. In other words, decision makers have an obligation to provide opportunities to people who have taken actions for which they deserve to be rewarded.

This principle can help to explain why we believe that people who plan to start a family should have the same chance of securing a job as others who do not when they demonstrate equal ability and ambition, even if starting a family requires employees to go on leave for extended periods and even if it increases the likelihood that employees will quit, thereby imposing costs on employers that might otherwise be avoided. While the employer's goal might be to recruit people who are likely to work diligently without interruption and who are likely to remain at the company indefinitely, selecting among applicants on this basis might cause an employer to disfavor applicants who deserve to be hired in light of their ability and ambition. Notably, would-be mothers, who are more likely than their peers to take time off or quit their jobs to start a family (due to entrenched gender norms around the division of parental responsibilities), should not be passed over in favor of others if they have all made the same investments in preparing themselves to apply for these positions. The principle of desert says that those who have cultivated the necessary skills to succeed on the job should all be assessed similarly, regardless of differences in the likelihood that applicants will need to take time off or give up their jobs.

Discussions of merit and desert also help to highlight that there can be quite different justifications for the constraints or demands that we might place on decision makers. In some cases, we might argue that people are simply morally

entitled to certain treatment. For example, we might say that it is wrong to hold people responsible for characteristics about themselves over which they have no control, even if doing so would be in the rational interest of a decision maker. Likewise, we might say that people are owed certain opportunities in light of their ability and ambition, even if the decision maker would prefer to judge people on a different basis. These are what philosophers call deontological arguments: moral reasons why some actions are preferable to others regardless of the consequences of these actions. We must discount the effects of bad luck and take merit into account because that is what fairness demands.

In contrast, we might argue that the way decision makers treat people should be dictated by the consequences of such actions. For example, we might say that merit-based decision making is justified on the grounds that allocating opportunities according to merit helps to advance the interest of society, not just the individual seeking a particular opportunity or the decision maker providing the opportunity. Hiring on the basis of ability and ambition may have the consequences of enhancing overall welfare if it means people who are particularly well prepared to undertake some activity are more likely to obtain the opportunity to do so. Merit-based decision making is thus justified because it puts individuals' talents to good use for society's collective benefit — not because any given individual is morally entitled to a particular opportunity in light of their merits. We might further argue that differential treatment on the basis of merit incentivizes and rewards productive investment that benefits all of society.

Of course, there are also consequentialist arguments in favor of interventions designed to uplift those who have experienced disadvantage or discrimination in the past. For example, society suffers overall when members of specific groups are denied the opportunity to realize their true potential because society forgoes the collective benefits that might be brought about by the contributions of such groups.

The cost of fairness

Different views of equality of opportunity — as well as the notions of merit and desert on which such views depend — allocate the responsibility and associated costs of dealing with unfairness quite differently. Notably, the middle view places the burden on individual decision makers and specific institutions regardless of the speed with which or the extent to which a person is able to realize their potential. For example, we might expect universities to incur some up front costs in admitting students from less privileged backgrounds because universities may have to invest additional resources in helping bring those students up to speed with their more privileged peers. This could take the form of providing classes over the summer leading up to the start of formal undergraduate programs. Or it could take the form of designing introductory courses without taking much background knowledge as a given, which might spend some time reviewing material that is familiar to students from better funded high schools, but perhaps new to those who come from less well-funded school districts. Universities might

even invest in programs that seek to limit the degree to which the inequalities that exist between students prior to enrolling in college *carry through* their college experiences. For example, universities might offer financial scholarships to poorer students with the goal of allowing them to avoid having to work in order to support themselves, thereby allowing these students to devote a similar amount of time to their studies as their more privileged peers. Such scholarships could also help to avoid saddling poorer students with significant debt, which might suppress future earnings and negatively influence career choices — burdens that richer students without significant debt need not navigate. Interventions along these lines blur the distinction between the middle and broad view of equality of opportunity because they seem targeted not at remedying some past unjust inequality but at preventing an unjust inequality from re-emerging. Chapter 8 will cover such interventions in greater depth.

Despite these efforts, universities may find that their investments in these kinds of interventions may take many years to pay off: students from less privileged backgrounds might trail their peers from more privileged backgrounds in the grades that they obtain over the course of their undergraduate careers, but ultimately achieve comparable success once they enter the labor market.

Likewise, employers who hire candidates that they recognize as having great potential, but also the need for additional support, might not be the employers who enjoy the payoff of such investments. Employees might take another job before the original employer feels that it has recouped its investment. This is an important aspect of the middle view of equality of opportunity because it highlights that it might not always be in the rational interest of decision makers to behave in these ways. (This might cut the other way as well, though: an unconstrained decision maker might discount someone who seems meritorious because the decision maker recognizes that the person has benefited from good luck — and is thus lacking in the ability or ambition that they are actually searching for.)

The middle view is thus not simply an argument that decision makers must attend to their long-term self interests; it is an argument that certain institutions are the right actors to incur some cost in the service of remedying inequality and injustice, even if there is no guarantee of obtaining a reward of at least equivalent value.

This contrasts with the broad view of equality of opportunity, where the government is understood to be the appropriate actor to facilitate the necessary redistribution to compensate for unjust disparities, likely through direct taxation and transfers. According to the broad view, the government — which is to say, everyone who pays taxes to the government — bears the burden to counteract the advantages that would otherwise be enjoyed by, for example, students from more privileged backgrounds. To the extent that interventions by employers or other institutions are necessary, the government should subsidize their efforts with tax money. In contrast, the middle view places this burden on specific decision makers to compensate for the disadvantages that people have already experienced.

This all suggests a number of difficult questions: To what extent should the burden for past discrimination fall on individual decision makers? On what times

scale should we attempt to correct the effects of historical injustice? And is it even possible to offset the cumulative result of the thousands of moments in which people treat each other unequally over the course of a lifetime? We'll return to these questions in Chapter 6 when we consider, from a legal and practical perspective, who we might view as best positioned to incur these costs.

Connecting statistical and moral notions of fairness

We now attempt to map some of the moral notions we've discussed so far in this chapter to the statistical criteria from Chapter 3. Of course, many of the concepts in this chapter, such as whether a decision subject has control over an attribute used for decision making, cannot be expressed in the statistical language of conditional probabilities and expectations. Further, even for notions that do seem to translate to statistical conditions, we reiterate our usual note of caution that statistical criteria alone cannot certify that a system is fair. As just one reason for this, the criteria in Chapter 3 are invariant to the application rates of different groups. For example, if 50% of loan applicants from a particular group decided not to apply for some reason, a classifier that satisfied independence/demographic parity before the change in application rates would still satisfy independence/demographic parity after the change. The same is true of sufficiency/calibration and separation/error rate parity. Yet, we wouldn't consider a bank, employer, or another institution to be fair if it discouraged applications from certain people or groups. This is related to Selbst et al.'s *framing trap*: a "failure to model the entire system over which [fairness] will be enforced".[22]

But we must also resist the opposite extreme, which is the view that statistical criteria have no normative content. We take the position that statistical criteria are one facet of what it means for a sociotechnical system to be fair and, combined with procedural protections, can help us achieve different moral goals.

Demographic parity

With those caveats out of the way, let's start with a relatively simple statistical criterion: demographic parity. It has a tenuous but discernible relationship to the broad view of equality of opportunity insofar as it aims to equalize outcomes. The high-level similarity between the two is the idea of proportional distribution of resources. But moral notions never map exactly to technical criteria. Let us look at the differences between them as a way of understanding the relationship.

The broad view of equality of opportunity is concerned with people's *life* outcomes (such as wealth) rather than discrete moments of decision making. Still, we may hope that imposing some notion of equality in decisions that affect the outcome of interest (such as jobs in the case of wealth) will lead to equality in the corresponding outcome. Empirically, however, it is far from clear that imposing equality in the short term will lead to equality in the long term. In fact, theoretical work has shown that this is not always the case.[23]

Further, equality of outcome — that is, enforcing equal life outcomes — ignores differences in ability and ambition between people that might reasonably justify differences in outcomes. This also happens to be the most common criticism of equality of outcome: it rules out the particular understanding of merit-based decision making that underpins the application of machine learning in many settings.

Even though this is often considered a fatal objection to equality of outcome, the criticism loses much of its force when applied to demographic parity. To try to justify demographic parity despite individual differences in ability and ambition, we acknowledge those differences but argue that these cancel out at the level of groups; thus, while decisions made about individuals can be attuned to the differences between them, we require the benefits and burdens of those decisions to be equally distributed among groups, on average.

But which groups should we pay attention to? As before, we pay special attention to group differences when we consider them especially likely to arise due to unjust historical conditions or to compound over time. These correspond to the axes along which society was historically and is currently stratified. In this view, we may care about equality of outcome not just for its own sake but also because inequality of outcome is a good indicator that there might be inequality of opportunity in the broad sense of the term.[24] In other words, certain inequalities in outcomes might not have arisen had there not been some past inequalities in opportunity.

There are many other gaps between demographic parity and equality of outcome. We'll mention just one more: not all decision subjects (and groups) may value the resource equally. Targeted ads may be helpful to wealthier individuals by informing them about things their money can buy, but prey upon the economic insecurities of poorer individuals (e.g. payday loans[25]). Policing may be helpful to some communities but put a burden on others, depending on the prejudices of police officers. In these cases, actual outcomes — benefits and harms — can be vastly different despite statistical parity in allocation.

Calibration

Recall from Chapter 3 that if group membership is redundantly encoded in the features, which is roughly true in sufficiently rich datasets, then calibration is a consequence of unconstrained supervised learning. Thus, it can be achieved without paying explicit attention to group membership. In other words, imposing calibration as a requirement is not much of an intervention.

Still, the notion has intuitive appeal: if a score is calibrated by group, then we know that a score value (say, 10% risk of default) indicates the same rate of positive outcomes (e.g., default rate) in all groups. By the same token, it has some diagnostic usefulness from a fairness perspective such as flagging 'irrational' discrimination. If the classifier explicitly encodes a preference for one group or a prejudice against another (or a human decision maker exercises such a preference or prejudice), the resulting distribution will not be calibrated by group.

Calibration can also be viewed as a sanity check for optimization. Precisely because calibration is implied by unconstrained optimization, we can detect optimization failures from violations of calibration. But that's all it is: a sanity check. A model can be egregiously inaccurate and still satisfy calibration. Indeed, a model with no discriminative power that always simply outputs the mean outcome of the population is perfectly calibrated. A model that is highly accurate for one group (optimal as defined in Chapter 3) while always predicting the mean for another group is also perfectly calibrated.

Calibration by group fits with a narrow view of equality of opportunity. Suppose that a decision maker uses only features deemed relevant, while group membership is deemed irrelevant. Then calibration by group says that the decision maker does not consider group membership beyond the extent to which it is encoded in task-relevant features. The decision may justify group differences in outcomes by appeal to differences in relevant features.

Some nontrivial normative justification is required for violating calibration in models used for decision making. We have discussed many such justifications, such as a belief that the risk arises partly from factors that the decision subject should not be held accountable for.

The similarity criterion

Let's return to the similarity criterion: treating similar people similarly. As we discussed, the normative substance of this notion largely comes down to what we mean by similar. One common view is to think of it as closeness with respect to features that relate to qualifications for the task at hand, interpreting features at face value.

To translate this to a technical notion, we can imagine defining a task-specific similarity function or metric between two feature vectors representing individuals. We can then insist that for any two individuals who are sufficiently similar, the decisions they receive be correspondingly similar. We call this the similarity criterion. This notion was made precise and analyzed by Cynthia Dwork, Moritz Hardt, Toniann Pitassi, et al.[16] Once we have a metric, we can solve a constrained optimization problem. The optimization objective is as usual (e.g., minimizing the difference between predicted and observed job performance) and the similarity criterion is the constraint.

We can illustrate this approach in the context of online behavioral advertising. Our discussion assumes that we view advertisements as allocating access to opportunity (e.g., through targeting job openings or credit offers). Ad networks collect demographic information about individuals, such as their browsing history, geographical location, and shopping behavior, and utilize this to assign a person to one of a few dozen segments. Segments have names such as "White Picket Fences," a market category with median household income of just over $50,000, aged 25 to 44 with kids, with some college education, etc. Individuals in a segment are considered similar for marketing purposes, and advertisers are allowed to target ads only at the level of segments and not individuals.

This reflects the narrow view of equality of opportunity. If two individuals differ only on dimensions that are deemed irrelevant to the advertiser's commercial interests, say religion, they will be in the same segment and thus are expected to see the same ads. On the other hand, if some people or social groups have had advantages throughout their lives that have enabled a certain income level, then the similarity criterion allows the benefits of those advantages to be reflected in the ads that they see.

Targeted advertising is a particularly good domain to apply these ideas. There is an intermediary — the ad network — that collapses feature vectors into categories (i.e., ad segments), and only exposes these categories to the advertisers, rather than directly allowing advertisers to target individuals. The ad network should construct the segments in such a way that members who are similar for advertising purposes must be in the same segment. For example, it would not be acceptable to include a segment corresponding to disability, because disability is not a relevant targeting criterion for the vast majority of types of ads. In domains other than targeted advertising, say college admissions, applying these ideas is more challenging. Absent an intermediary like the ad network, it is up to each decision maker to provide transparency into their similarity metric.

This narrow interpretation of the similarity criterion relates to other formal definitions of individual fairness, such as, the notion of *meritocratic fairness* in the context of bandit learning.[26] The normative content of the similarity criterion, however, extends beyond the narrow view of equality of opportunity if we broaden the principles from which we construct a similarity metric. For example, the notion of similarity might explicitly adjust features based on consideration of past injustice and disadvantage. We might agree at the outset that an SAT test score of 1200 under certain circumstances corresponds to a score of 1400 under more favorable background conditions.

Comparisons such as these are closely related to Roemer's formal definition of equality of opportunity.[21] Roemer envisions a partition of the population into *types* based on "easily observable and nonmaniupulable" features that relate to "differential circumstances of individuals for which we believe they should not be held accountable". The formal definition then compares individuals who expend the same quantile of *effort* relative to their type.

Randomization, thresholding, and fairness

If we think about applying the similarity criterion to a task like hiring, we run into another problem: pairs of candidates who are extremely similar may fall on opposite sides of the score threshold, because we have to draw the line somewhere. This would violate the similarity criterion. One way to overcome this is to insist that the classifier be randomized.

Randomization sometimes offends deeply held moral intuitions, especially in domains such as criminal justice, by conjuring the specter of decisions made based on a coin flip. But there are several reasons why randomization may not just be acceptable but necessary for fairness in some cases (in addition to the fact that it

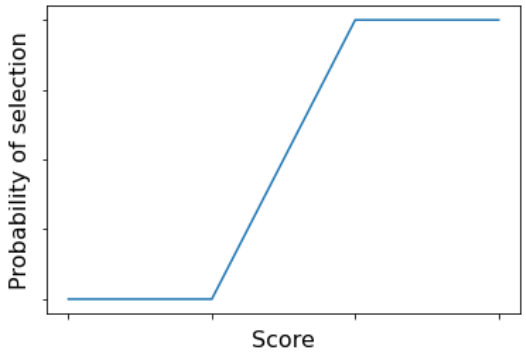

Figure 1: A randomized classifier. Only randomized classifiers can satisfy the similarity criterion. Two similar individuals would have similar scores and thus similar probabilities of selection.

allows us to treat similar people similarly, at least in a probabilistic sense). In fact, Ronen Perry and Tal Zarsky present numerous examples of cases where the law requires that consequential decisions be based on lotteries.[27]

To understand the justification for randomized decision making, we must recognize that precisely controlled and purposeful randomness is not the same as arbitrariness or capriciousness. Suppose these three conditions hold: a resource to be allocated is indivisible, there are fewer units of it than claimants, and there is nothing that entitles one claimant to the resource any more or any less than other claimants. Then randomization may be the *only* egalitarian way to break the tie. This is the principle behind lotteries for allocation of low-rent public housing and immigration visas. The same principle applies to burdens rather than valued resources, as seen in the random selection of people for jury duty or tax audits.[27]

But the whole point of employing machine learning is that there does exist a way to rank the claims of the applicants, so the scenarios of interest to us are more complicated than the above examples. The complication is that there is a conflict between the goals of treating similar people similarly (which requires randomization) and minimizing unpredictability in the decision (which requires avoiding randomization).

One critical distinction that affects the legitimacy of randomization is whether there are equivalent opportunities for which an applicant might be eligible. This is generally true in the case of hiring or lending, and not true in the case of pre-trial detention. Randomization is more justifiable in the former case because it avoids the problem of an applicant perpetually falling just short of the selection threshold. If randomization is employed, a reasonably qualified but not stellar job applicant might have to apply to several jobs, but will eventually land one.[28]

Another way to avoid the problem of similar candidates falling on opposite sides of a cutoff is to redesign the system so that decisions are not binary. This is again easier for some institutions than others. A lender can account for different levels of risk by tailoring the interest rate for a loan rather than rejecting the loan

altogether. In contrast, a binary notion of determination of guilt is built into the criminal justice system.[6] This is not easy to change. Note that determinations of guilt are not predictions; they are meant to reflect some binary ground truth and the goal of the criminal justice system is to uncover it.

The normative underpinnings of error rate parity

Of the three main families of statistical criteria in Chapter 3, we have discussed independence / demographic parity and sufficiency / calibration, leaving separation / error rate parity. Error rate parity is the hardest criterion to rigorously connect to any moral notion. At the same time, it is undeniable that it taps into some widely held fairness intuition. ProPublica's study of the COMPAS criminal risk prediction system was so powerful because of the finding that Black defendants had "twice the false positive rate" of White defendants.[29]

But there is no straightforward justification for this intuition, which has led to error rate parity becoming a topic of fierce debate.[30,31] Building on this scholarship, we provide our view of why we should care about error rate parity.

We'll assume a prediction-based resource allocation problem such as lending that has a substantial degree of inherent uncertainty with respect to the predictability of outcomes. In contrast, error rate disparity often comes up in perception problems like facial recognition or language detection where there is little or no inherent uncertainty.[32,33] The crucial normative difference is that in face recognition, language detection, and similar applications, there is no notion of a difference in qualification between individuals that could potentially justify dissimilar treatment. Thus, assuming that misclassification imposes a cost on the subject, it is much more straightforward to justify why unequal error rates are problematic.

Another observation to set the stage: the moral significance of error rate is asymmetric. One type of error, roughly speaking, corresponds to unjust denial (of freedom or opportunity) and the other corresponds to overly lenient treatment. In most domains, the first type is much more significant as a normative matter than the second. For example, in the context of bail decisions, it is primarily the disparity in the rates of pretrial detention of non-recidivists that's worrisome, rather than disparities in the rates of pretrial release of recidivists. While it is true that the release of would-be recidivists has a cost in the form of a threat to public safety, that cost depends on the *total* error and not the distribution of that error between groups. Thus, it is not necessarily meaningful to simply compare error rates between groups.

Error rate parity and the middle view of equality of opportunity

Recall that the middle view of equality of opportunity takes into account historical and present social conditions that may affect why people's qualifications may differ.

[6] That said, there have been proposals to envision an alternate system where the degree of punishment is calibrated to the strength of the evidence. Schauer, Frederick. *Profiles, Probabilities, and Stereotypes*. Cambridge, MA: Harvard University Press, 2006.

To understand a decision making system with respect to the middle view, it is critical to know if the effects of the decisions might themselves perpetuate these conditions in society.

Unfortunately, this is hard to do with the data available at the moment of decision making, especially if the features (that encode decision subjects' qualifications) are not available. One thing we can do even without the features is to look at differences in base rates (i.e., rates at which different groups achieve desired outcomes, such as loan repayment or job success). If the base rates are significantly different — and if we assume that individual differences in ability and ambition cancel out at the level of groups — it suggests that people's qualifications may differ due to circumstances beyond the individual.

But base rates alone don't shed light on whether the classifier might perpetuate existing inequalities. For this analysis, what's important is whether the classifier imposes an unequal *burden* on different groups. There are many reasonable ways to measure the burden, but since we consider one type of error — mistakenly classifying someone as undeserving or high-risk — to be especially egregious, we can consider the rate of such misclassification among members of a group as a proxy for the burden placed on that group. This is especially true when we consider the possibility of spillover effects: for example, denying pretrial release has effects on defendants' families and communities.

When a group is burdened by disproportionately high error rates, it suggests that the system may perpetuate cycles of inequality. Indeed, Aziz Huq argues that for this reason, the criminal justice system entrenches racial stratification, and this is the primary racial inequity in algorithmic criminal justice.[34] To be clear, the effect of institutions on communities is an empirical and causal question that cannot be boiled down to error rates, but given the limitations of observational data available in typical decision making scenarios, error rates are a starting point for investigating this question. This yields a distinct reason why error rates carry some moral significance. But note a finding of error rate disparity, by itself, doesn't suggest any particular intervention.

What to do about error rate disparity

Collecting more data and investing in improving classification technology is one way to potentially mitigate error rate disparity. Normally, we give significant deference to the decision maker on the tradeoff between data collection cost and model accuracy. This deference, especially in private sector applications, is based on the idea that the interests of the decision maker and decision subjects are generally aligned. For example, we defer to lenders on how accurate their predictions should be. If, instead, lenders were required to be highly accurate in their predictions, they might only lend in the safest of cases, depriving many people of the ability to obtain loans, or they might go to great lengths to collect data about borrowers, raising the cost of operating the system and pushing some of that cost to the borrowers.

The argument above only considers total welfare and not how the benefits and

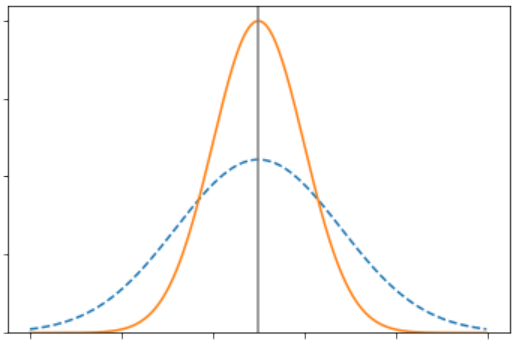

Figure 2: Probability density of risk scores for two groups, and a classification threshold. Throughout the illustrations in this section, we assume that the score is perfectly calibrated. The group shown with a solid line has a higher error rate. Intuitively this is because the probability mass is more concentrated (i.e., the score function is worse at distinguishing among members of this group). Collecting more data would potentially bring the solid curve closer to the dashed curve, mitigating the error rate disparity.

costs are distributed among people and groups. When we introduce distributional considerations, there are many scenarios where it is justifiable to lower the deference to decision makers, and the presence of error rate disparity is one such scenario. In this case, requiring the decision maker to mitigate error rate disparity can be seen as asking them to bear some of the cost that's being pushed onto some individuals and groups.

While improving the overall accuracy of the classifier may close the disparity in some cases, in other cases it may leave the disparity unchanged or even worsen it. Accuracy is bounded by that of the optimal classifier, and recall that the optimal classifier doesn't necessarily satisfy error rate parity. As a concrete example, assume that loan defaults primarily arise due to unexpected job loss, one group of loan applicants holds more precarious jobs that are at a greater risk of layoff, and layoffs are not predictable at decision time. In this scenario, improvements in data collection and classification will not yield error rate parity.

Faced with this intrinsic limitation, it may be tempting to perform an adjustment step that achieves error rate parity, such as different risk thresholds for different groups. One way to do this would be without making anyone worse off compared to an unconstrained classifier. For example, a lender could use a more lenient risk threshold for one group to lower its error rate. This would violate the narrow view of equality of opportunity, as people from different groups with the same risk score may be treated differently. Whether the intervention is still justified is a difficult normative question that lacks a uniform answer.

In other situations, even this might not be possible. For example, the intervention may increase the lender's risk so much that it goes out of business.

In fact, if base rates are so different that we expect large disparities in error rates that cannot be mitigated by interventions like data collection, then it suggests

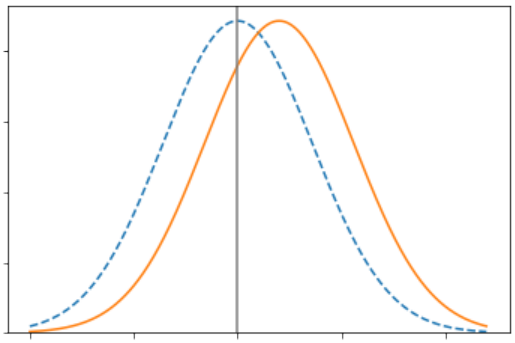

Figure 3: Probability density of risk scores for two groups, and a classification threshold. Again the solid group has a higher error rate—specifically, a higher false positive rate, where false positives are people incorrectly classified as high risk. But this time it is because the solid group has a higher base rate (the curve is shifted to the right compared to the dashed group). Collecting more data is unlikely to mitigate the error rate disparity.

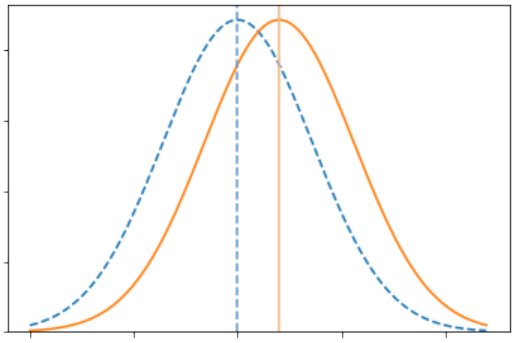

Figure 4: Probability density of risk scores for two groups, and two different classification thresholds resulting in equal error rates.

that the use of predictive decision making is itself problematic, and perhaps we should scrap the system or apply more fundamental interventions.

In summary, error rate parity lacks a direct relationship to any single normative principle. But it captures something about both the narrow and the middle views of equality of opportunity. It is also a way to incentivize decision makers to invest in fairness and to question the appropriateness of predictive decision making.

Alternatives for realizing the middle view of equality of opportunity

We've discussed how error rate parity bears some relationship to the middle view of equality of opportunity. But there are many other possible interventions that decision makers might adopt to try to realize the middle view of equality of opportunity that do not map onto any of the criteria discussed in Chapter 3. The middle view is an inherently fuzzy notion, leaving a lot of room to decide the extent to which we want to discount people's apparent differences and the manner in which to do so. Here are a few other ways in which we can try to operationalize it. Unsurprisingly, all of these violate the narrow view of equality of opportunity.

Decision makers could reconsider the goals that they are pursuing such that the decision-making process that seeks to meet these goals generates less disparate outcomes. For example, employers might choose a different target variable that is perceived to be an equally good proxy for their goal, but whose accurate prediction leads to a less significant disparity in outcomes for different groups.[35]

They might explore whether it is possible to train alternative models with a similar degree of accuracy as their original model, but which produces smaller disparities in the rates at which members of different groups achieve the desired outcome or are subject to erroneous assessment.[36] Empirically, this appears to be possible in many cases, including for particularly high-stakes decision making.[37]

They might sacrifice a good deal of apparent accuracy on the belief that there is serious measurement error and that people from some groups are actually far more qualified than they might appear (we assume that it is not possible to explicitly correct the measurement error and that group membership is not sufficiently redundantly encoded in the features, preventing the optimal classifier from automatically accounting for measurement error).[38]

Finally, they could forgo some of the benefits they might have achieved under the original decision-making process so as to provide important benefits to the groups that have been subject to past mistreatment. To do so, they might treat members of certain groups counterfactually, as if they hadn't experienced the injustice that makes them less qualified by the time of decision-making.

Summary

Fairness is most often conceptualized as equality of opportunity. But in this chapter, we've seen that there are a variety of ways to understand equality of opportunity. The differences among them are at the heart of why fairness is such a contested

topic. All three views can be seen in contemporary political debates. The narrow view aligns with what is often meant by the term meritocracy. The middle view drives Diversity, Equity, and Inclusion (DEI) efforts at many workplaces. The broad view is too sweeping to find much support for a full-throated implementation, but the ideas behind it come up in debates around topics such as reparations.[39]

The views differ along many axes, including what they seek to achieve; how they understand the causes of current differences between groups (and whether they seek to understand them at all); and how to distribute the cost of uplifting historically disadvantaged groups.

Table 2: Views of equality of opportunity and their formal relatives

	Goal	Related formal criteria
Narrow view	Ensure that people who are similarly qualified for an opportunity have similar chances of obtaining it	Similarity criterion, meritocratic fairness, calibration by group
Middle view	Discount differences due to past injustice that accounts for current differences in qualifications	Similarity criterion, Roemer's formal equality of opportunity, error rate parity
Broad view	Ensure people of equal ability and ambition are able to realize their potential equally well	Demographic parity

In the latter part of the chapter, we attempted to connect these moral notions to the statistical criteria from Chapter 3. Loose connections emerged through this exercise, but, ultimately, none of the statistical criteria are strongly anchored in normative foundations.

But even these rough similarities illustrate one important point about the impossibility results from Chapter 3. The impossibility results aren't some kind of artifact of statistical decision making; they simply reveal moral dilemmas. Once we recognize the underlying moral difficulties, these mathematical tensions seem much less surprising.

For example, an approach that makes accurate predictions based on people's currently observable attributes, and then makes decisions based on those predictions (calibration) won't result in equality of outcomes (independence) as long as different groups have different qualifications on average. Similarly, its results also differ from an approach that is willing to treat seemingly similar people differently in order to attempt to equalize the burden on different groups (error rate parity). The approaches also differ in the extent to which measurement errors are seen as the responsibility of the decision maker, and who should bear the costs of fairness interventions.

One reason the normative foundations of statistical fairness criteria are shaky is that conditional independence doesn't give us a vocabulary to reason about

the causes of disparities between groups or the effects of interventions. We will attempt to address these limitations in the next chapter.

Bibliography

[1] Hellman, Deborah. 2007. *When Is Discrimination Wrong?* Harvard University Press, Cambridge, MA.

[2] Lippert-Rasmussen, Kasper. 2013. *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*. Oxford University Press, Oxford, UK.

[3] Sunstein, Cass R. 1994. "The Anticaste Principle". *Michigan Law Review*, 92(8):2410.

[4] Singer, Peter. 1978. "Is racial discrimination arbitrary?" *Philosophia*, 8(2-3):185–203.

[5] Schauer, Frederick. 2006. *Profiles, probabilities, and stereotypes*. Harvard University Press, Cambridge, MA.

[6] Alexander, Larry. 1992. "What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies". *University of Pennsylvania Law Review*, 141(1):149.

[7] Arneson, Richard J.. 2006. "What Is Wrongful Discrimination ". *San Diego Law Review*, 43(4):775–808.

[8] Eidelson, Benjamin. 2015. *Discrimination and Disrespect*. Oxford University Press, New York, NY.

[9] Balkin, J M. 1997. "The Constitution of Status". *The Yale Law Journal*, 106(8):2313–2374.

[10] Hoffman, Sharona. 2011. "The Importance of Immutability in Employment Discrimination Law". *William & Mary Law Review*, 52(5):1483–1546.

[11] Clarke, Jessica A.. 2015. "Against Immutability". *Yale Law Journal*, 125(1):1–102.

[12] Hellman, Deborah. 2020. "Indirect Discrimination and the Duty to Avoid Compounding Injustice". In Hugh Collins and Tarunabh Khaitan (editors), *Foundations of Indirect Discrimination Law*. Bloomsbury Publishing, London.

[13] Schauer, Frederick. 2017. "Statistical (and Non-Statistical) Discrimination". In Kasper Lippert-Rasmussen (editor), *The Routledge Handbook of the Ethics of Discrimination*. Routledge, New York, NY.

[14] Hellman, Deborah. 2017. "Discrimination and Social Meaning". In Kasper Lippert-Rasmussen (editor), *The Routledge Handbook of the Ethics of Discrimination*. Routledge, New York, NY.

[15] Prince, Anya E.R. and Schwarcz, Daniel. 2020. "Proxy Discrimination in the Age of Artificial Intelligence and Big Data". *Iowa Law Review*, 105(3):1257–1318.

[16] Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard. 2012. "Fairness through awareness". In *Proc. 3rd ITCS*, pages 214–226.

[17] Anderson, Elizabeth S. 1999. "What Is the Point of Equality?" *Ethics*, 109(2):287–337.

[18] Rawls, John. 1998. *A Theory of Justice*. Harvard University Press, Cambridge, MA.

[19] Arneson, Richard. 2018. "Four Conceptions of Equal Opportunity". *The Economic Journal*, 128(612):F152–F173.

[20] Fishkin, Joseph. 2013. *Bottlenecks*. Oxford University Press, New York, NY.

[21] Roemer, John. 2000. *Equality of Opportunity*. Harvard University Press, Cambridge, MA.

[22] Selbst, Andrew D., Boyd, Danah, Friedler, Sorelle A., Venkatasubramanian, Suresh, and Vertesi, Janet. 2019. "Fairness and Abstraction in Sociotechnical Systems". In *Conference on Fairness, Accountability, and Transparency*, pages 59–68.

[23] Liu, Lydia T., Dean, Sarah, Rolf, Esther, Simchowitz, Max, and Hardt, Moritz. 2017. "Delayed Impact of Fair Machine Learning". In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3150–3158.

[24] Phillips, Anne. 2004. "Defending Equality of Outcome". *Journal of Political Philosophy*, 12(1):1–19.

[25] Rieke, Aaron and Koepke, Logan. 2015. "Led Astray: Online Lead Generation and Payday Loans". Technical report, Upturn, Washington, DC.

[26] Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie H, and Roth, Aaron. 2016. "Fairness in learning: Classic and contextual bandits". In *Advances in Neural Information Processing Systems*, pages 325–333.

[27] Perry, Ronen and Zarsky, Tal. 2015. ""May the Odds Be Ever in Your Favor": Lotteries in Law". *Alabama Law Review*, 66(5):1035–1098.

[28] Creel, Kathleen and Hellman, Deborah. 2022. "The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems". *Canadian Journal of Philosophy*, pages 1–18.

[29] Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. 2016. "Machine bias". *ProPublica*.

[30] Matwin, Stan, Yu, Shipeng, Farooq, Faisal, Corbett-Davies, Sam, Pierson, Emma, Feller, Avi, Goel, Sharad, and Huq, Aziz. 2017. "Algorithmic Decision Making and the Cost of Fairness". *International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

[31] Hellman, Deborah. 2022. "Measuring Algorithmic Fairness". *Virginia Law Review*, 106(4):811–866.

[32] Krishnapriya, K.S, Vangara, Kushal, King, Michael C., Albiero, Vitor, and Bowyer, Kevin. 2019. "Characterizing the Variability in Face Recognition Accuracy Relative to Race". volume 00 of *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2278–2285.

[33] Blodgett, Su Lin, Green, Lisa, and O'Connor, Brendan. 2016. "Demographic Dialectal Variation in Social Media: A Case Study of African-American English". In *Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

[34] Huq, Aziz Z. 2018. "Racial equity in algorithmic criminal justice". *Duke LJ*, 68:1043.

[35] Passi, Samir and Barocas, Solon. 2019. "Problem formulation and fairness". In *Conference on Fairness, Accountability, and Transparency*, pages 39–48.

[36] Black, Emily, Raghavan, Manish, and Barocas, Solon. 2022. "Model Multiplicity: Opportunities, Concerns, and Solutions". In *Conference on Fairness, Accountability, and Transparency*, pages 850–863.

[37] Rodolfa, Kit T., Lamba, Hemank, and Ghani, Rayid. 2021. "Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy". *Nature Machine Intelligence*, 3(10):896–904.

[38] Friedler, Sorelle A., Scheidegger, Carlos, and Venkatasubramanian, Suresh. 2021. "The (Im)possibility of fairness". *Communications of the ACM*, 64(4):136–143.

[39] Hannah-Jones, Nikole. 2020. "What is owed". *The New York Times*.