2

When is automated decision making legitimate?

These three scenarios have something in common:

- A student is proud of the creative essay she wrote for a standardized test. She receives a perfect score, but is disappointed to learn that the test had in fact been graded by a computer.
- A defendant finds that a criminal risk prediction system categorized him as high risk for failure to appear in court, based on the behavior of others like him, despite his having every intention of appearing in court on the appointed date.
- An automated system locked out a social media user for violating the platform's policy on acceptable behavior. The user insists that they did nothing wrong, but the platform won't provide further details nor any appeal process.

All of these are automated decision-making or decision support systems that likley feel unfair or unjust. Yet this is a sense of unfairness that is distinct from what we talked about in the first chapter (and which we will return to in the next chapter). It is not about the relative treatment of different groups. Instead, what these questions are about is *legitimacy* — whether it is fair to deploy such a system at all in a given scenario. That question, in turn, affects the legitimacy of the organization deploying it.

Most institutions need legitimacy to be able to function effectively. People have to believe that the institution is broadly aligned with social values. The reason for this is relatively clear in the case of public institutions such as the government, or schools, which are directly or indirectly accountable to the public.

It is less clear why private firms need legitimacy. One answer is that the more power a firm has over individuals, the more the exercise of that power needs to be perceived as legitimate. And decision making about people involves exercising power over them, so it is important to ensure legitimacy. Otherwise, people will find various ways to resist, notably through law. A loss of legitimacy might also hurt a firm's ability to compete in the market.

Questions about firms' legitimacy have repeatedly come up in the digital technology industry. For example, ride sharing firms have faced such questions, leading to activism, litigation, and regulation. Firms whose business models rely on personal data, especially covertly collected data, have also undergone crises of perception. In addition to legal responses, such firms have seen competitors capitalize on their lax privacy practices. For instance, Apple made it harder for Facebook to track users on iOS, putting a dent in its revenue.[1] This move

enjoyed public support despite Facebook's vociferous protests, arguably because the underlying business model had lost legitimacy.

For these reasons, a book on fairness is incomplete without a discussion of legitimacy. Moreover, the legitimacy question should precede other fairness questions. Debating distributive justice in the context of a fundamentally unjust institution is at best a waste of time, and at worst helps prop up the institution's legitimacy, and is thus counterproductive. For example, improving facial analysis technology to decrease the disparity in error rates between racial groups is not a useful response to concerns about the use of such technologies for oppressive purposes.[2]

Discussions of legitimacy have been largely overshadowed by discussions of bias and discrimination in the fairness discourse. Often, advocates have chosen to focus on distributional considerations as a way of attacking legitimacy, since it tends to be easier argument to make. But this can backfire, as many firms have co-opted fairness discourse, and find it relatively easy to ensure parity in the decisions between demographic groups without addressing the legitimacy concerns.[3]

This chapter is all about legitimacy: whether it is morally justifiable to use machine learning or automated methods at all in a given scenario.

Although we have stressed the overriding importance of legitimacy, readers interested in distributive questions may skip to Chapter 3 for a technical treatment or to Chapter 4 for a normative account; those chapters, Chapter 3 in particular, do not directly build on this one.

Machine learning is not a replacement for human decision making

Machine learning plays an important role in decisions that allocate resources and opportunities that are critical to people's life chances. The stakes are clearly high. But people have been making high stakes decisions about each other for a long time, and those decisions seem to be subject to far less critical examination. Here's a strawman view: decisions based on machine learning are analogous to decision making by humans, and so machine learning doesn't warrant special concern. While it's true that machine learning models might be difficult for people to understand, humans are black boxes, too. And while there can be systematic bias in machine learning models, they are often demonstrably less biased than humans.

We reject this analogy of machine learning to human decision making. By understanding why it fails and which analogies are more appropriate, we'll develop a better appreciation for what makes machine learning uniquely dangerous as a way of making high-stakes decisions.

While machine learning is sometimes used to automate the tasks performed inside a human's head, many of the high-stakes decisions that are the focus of the work on fairness and machine learning are those that have been traditionally performed by *bureaucracies*. For example, hiring, credit, and admissions decisions are rarely left to one person to make on their own as they see fit. Instead, these decisions are guided by formal rules and procedures, involving many actors with

prescribed roles and responsibilities. Bureaucracy arose in part as a response to the subjectivity, arbitrariness, and inconsistency of human decision making; its institutionalized rules and procedures aim to minimize the effects of humans' frailties as individual decision makers.[4]

Of course, bureaucracies aren't perfect. The very term bureaucracy tends to have a negative connotation — a needlessly convoluted process that is difficult or impossible to navigate. And despite their overly formalistic (one might say cold) approach to decision making, bureaucracies rarely succeed in fully disciplining the individual decision makers that occupy their ranks. Bureaucracies risk being equally capricious and inscrutable as humans, but far more dehumanizing.[4]

That's why bureaucracies often incorporate procedural protections: mechanisms that ensure that decisions are made transparently, on the basis of the right and relevant information, and with the opportunity for challenge and correction. Once we realize that machine learning is being used to automate bureaucratic rather than individual decisions, asserting that humans don't need to — or simply cannot — account for their everyday decisions does not excuse machine learning from these expectations. As Katherine Strandburg has argued, "[r]eason giving is a core requirement in conventional decision systems *precisely because* human decision makers are inscrutable and prone to bias and error, not because of any expectation that they will, or even can, provide accurate and detailed descriptions of their thought processes".[5]

In analogizing machine learning to bureaucratic — rather than individual — decision making, we can better appreciate the source of some of the concerns about machine learning. When it is used in high-stakes domains, it undermines the kinds of protections that we often put in place to ensure that bureaucracies are engaged in well-executed and well-justified decision making.

Bureaucracy as a bulwark against arbitrary decision making

The kind of problematic decision making that bureaucracies protect against can be called *arbitrary* decision making. Kathleen Creel and Deborah Hellman have usefully distinguished betweeen two flavors of arbitrariness.[6] First, arbitrariness might refer to decisions made on an inconsistent or ad hoc basis. Second, arbitrariness might refer to the basis for decision making lacking reasoning, even if the decisions are made consistently on that basis. This first view of arbitrariness is principally concerned with procedural regularity:[7] whether a decision making scheme is executed consistently and correctly. Worries about arbitrariness, in this case, are really worries about whether the rules governing important decisions are fixed in advance and applied appropriately, with the goal of reducing decision makers' capacity to make decisions in a haphazard manner.

When decision making is arbitrary in this sense of the term, individuals may find that they are subject to different decision-making schemes and receive different decisions simply because they happen to go through the decision-making process at different times. Not only might the decision-making scheme change over time;

human decision makers might be inconsistent in how they apply these schemes as they make their way through different cases. The latter could be true of one individual decision maker whose behavior is inconsistent over time, but it could also be true if the decision-making process allocates cases to different individuals who are individually consistent, but differ from one another. Thus, even two people who are identical when it comes to the decision criteria may receive different decisions, violating the expectation that similar people should be treated similarly when it comes to high-stakes decisions.

This principle is premised on the belief that people are entitled to similar decisions unless there are reasons to treat them differently (we'll soon address what determines if these are *good* reasons). For especially consequential decisions, people may have good reason to wonder why someone who resembled them received the desired outcome from the decision-making process while they did not.

Inconsistency is also problematic when it prevents people from developing effective life plans based on expectations about the decision-making systems they must navigate in order to obtain desirable resources and opportunities.[6] Thus, inconsistent decision making is unjust both because it might result in unjustified differential treatment of similar individuals and also because it is a threat to individual autonomy by preventing people from making effective decisions about how best to pursue their life goals.

The second view of arbitrariness is getting at a deeper concern: are there good reasons — or any reasons — why the decision-making scheme looks the way that it does? For example, if a coach picks a track team based on the color of runners' sneakers, but does so consistently, it is still arbitrary because the criterion lacks a valid basis. It does not help advance the decision maker's goals (e.g., assembling a team of runners that will win the upcoming meet).

Arbitrariness, from this perspective, is problematic because it undermines a bedrock justification for the chosen decision-making scheme: that it actually helps to advance the goals of the decision maker. If the decision-making scheme does nothing to serve these goals, then there is no justified reason to have settled on that decision-making scheme — and to treat people accordingly. When desirable resources and opportunities are allocated arbitrarily, it needlessly subjects individuals to different decisions, despite the fact that all individuals may have equal interest in these resources and opportunities.

In the context of government decision making, there is often a legal requirement that there be a *rational* basis for decision making — that is, that there be good reasons for making decisions in the way that they are.[6] Rules that do not help the government achieve its stated policy goals run afoul of the principles of due process. This could be either because the rules were chosen arbitrarily or because of some evident fault with the reasoning behind these rules. These requirements stem from the fact that the government has a monopoly over certain highly consequential decisions, leaving people with no opportunity to seek recourse by trying their case with another decision maker.

There is no corresponding legal obligation when the decision makers are private actors, as Creel and Hellman point out. Companies are often free to make poorly

reasoned — even completely arbitrary — decisions. In theory, decision-making schemes that seem to do nothing to advance private actors' goals should be pushed out of the market by competing schemes that are more effective.[6]

Despite this, we often expect that decisions of major consequence, even when they are made by private actors, are made for good reasons. We are not likely to tolerate employers, lenders, or admission officers that make decisions about applicants by flipping a coin or according to the color of applicants' sneakers. Why might this be?

Arbitrary decision making fails to respect the gravity of these decisions and shows a lack of respect for the people subject to them. Even if we accept that we cannot dictate the goals of institutions, we still object to giving them complete freedom to treat people however they like. When the stakes are sufficiently high, decision makers bear some burden for justifying their decision-making schemes out of respect for the interests of people affected by these decisions. The fact that people might try their luck with other decision makers in the same domain (e.g., another employer, lender, or admission officer) may do little to modulate these expectations.

Three Forms of Automation

To recap our earlier discussion, automation might undermine important procedural protections in bureaucratic decision making. But what, exactly, does machine learning help to automate? It turns out that there are three different types of automation.

The first kind of automation involves taking decision-making rules that have been set down by hand (e.g., worked out through a traditional policy-making process) and translating these into software, with the goal of automating their application to particular cases.[8] For example, many government agencies follow this approach when they adopt software to automate benefits eligibility determinations in accordance with pre-existing policies. Likewise, employers follow this approach when they identify certain minimum qualifications for a job and develop software to automatically reject applicants that do not possess them. In both of these cases, the rules are still set by humans, but their application is automated by a computer; machine learning has no obvious role here.

But what about cases where human decision makers have primarily relied on informal judgment rather than formally specified rules? This is where the second kind of automation comes in. It uses machine learning to figure out how to replicate the informal judgements of humans. Having automatically discovered a decision-making scheme that produces the same decisions as humans have made in the past, it then implements this scheme in software to replace the humans who had been making these decisions. The student whose creative essay was subject to computerized assessment, described in the opening of this chapter, is an example of just such an approach: the software in this case seeks to replicate the subjective evaluations of human graders.

The final kind of automation is quite different from the first two. It does not rely on an existing bureaucratic decision making scheme or human judgment. Instead, it involves learning decision-making rules from data. It uses a computer to uncover patterns in a dataset that predict an outcome or property of policy interest — and then bases decisions on those predictions. Note that such rules could be applied either manually (by humans) or automatically (through software). The relevant point of automation, in this case, is in the process of developing the rules, not necessarily applying them. For example, these could be rules that instruct police to patrol certain areas, given predictions about the likely incidence of crime based on past observations of crime. Or they could be rules that suggest that lenders grant credit to certain applicants, given the repayment histories of past recipients like them. Machine learning — and other statistical techniques — are crucial to this form of automation.

As we'll see over the next three sections, each type of automation raises its own unique concerns.

Automating Pre-Existing Decision-Making Rules

In many respects, the first form of automation — translating pre-existing rules into software so that decisions can be executed automatically — is a direct response to arbitrariness as inconsistency. Automation helps ensure consistency in decision making because it requires that the scheme for making decisions be fixed. It also means that the scheme is applied the same way every time.

And yet, many things can go wrong. Danielle Citron offers a compelling account of the dangers of automating decision-making rules established via a deliberative policy-making or rule-making process.[8] Automating the execution of a pre-existing decision-making scheme requires translating such a scheme into code. Programmers might make errors in that process, leading to automated decisions that diverge from the policy that the software is meant to execute. Another problem is that the policy that programmers are tasked with automating may be insufficiently explicit or precise; in the face of such ambiguity, programmers might take it upon themselves to make their own judgment calls, effectively usurping the authority to define policy. And at the most basic level, software may be buggy. For example, hundreds of British postmasters were convicted for theft or fraud over a twenty year period based on flawed software in what has been called the biggest miscarriage of justice in British history.[9]

Automating decision making can also be problematic when it completely stamps out any room for discretion. While human discretion presents its own issues, as described above, it can be useful when it is difficult or impossible to fully specify how decisions should be made in accordance with the goals and principles of the institution.[10] Automation requires that an institution determine in advance all of the criteria that a decision-making scheme will take into account; there is no room to consider the relevance of additional details that might not have been considered or anticipated at the time that the software was developed.

Automated decision-making is thus likely to be much more brittle than decision-

making that involves manual review because it limits the opportunity for decision subjects to introduce information into the decision-making process. People are confined to providing evidence that corresponds to a pre-established field in the software. Such constraints can result in absurd situations in which the strict application of decision-making rules leads to outcomes that are directly counter to the goals behind these rules. New evidence that would immediately reverse the assessment of a human decision maker may have no place in automated decision making.[11] For example, in an automated system to assess people with illnesses to determine eligibility for a state-provided caregiver, one field asked if there were any foot problems. An assessor visited a certain person and filled out the field to indicate that they didn't have any problems — because they were an amputee.[12]

Discretion is valuable in these cases because humans are often able to reflect on the relevance of additional information to the decision at hand and the underlying goal that such decisions are meant to serve. In effect, human review leaves room to expand the criteria under consideration and to reflect on when the mechanical application of the rules fails to serve their intended purpose.[13,11]

These same constraints can also restrict people's ability to point out errors or to challenge the ultimate decision.[14] When interacting with a loan officer, a person could point out that their credit file contains erroneous information. When applying for a loan via an automated process, they might have no equivalent opportunity. Or perhaps a person recognizes that the rules dictating their eligibility for government benefits have been applied incorrectly. When caseworkers are replaced by software, people subject to these decisions may have no means to raise justified objections.[15]

Finally, automation runs the serious risk of limiting accountability and exacerbating the dehumanizing effects of dealing with bureaucracies. Automation can make it difficult to identify the agent responsible for a decision; software often has the effect of dispersing the locus of accountability because the decision seems to be made by no one.[16] People may have more effective means of disputing decisions and contesting the decision-making scheme when decision-making is vested in identifiable people. Likewise, automation's ability to remove humans from the decision-making process may contribute to people's sense that an institution does not view them as worthy of the respect that would grant them an opportunity to make legitimate corrections, introduce additional relevant information, or describe mitigating circumstances.[17] This is precisely the problem highlighted by the opening example of a social media user who had been kicked off a platform without explanation or opportunity for appeal.

We've highlighted many normative concerns that arise from simply automating the application of a pre-existing decision-making scheme. While many of these issues are commonly attributed to the adoption of machine learning, none of them originate from the use of machine learning specifically. Long-standing efforts to automate decision-making with traditional software pose many dangers of their own. The fact that machine learning is not the exclusive cause of these types of problems is no reason to take them any less seriously, but effective responses to these problems requires that we be clear about their origins.

Learning Decision-Making Rules from Data on Past Decisions in order to Automate Them

Decision makers might have a pre-existing but informal process for making decisions which they might like to automate. In this case, machine learning (or other statistical techniques) might be employed to "predict" how a human would make a decision, given certain criteria. The goal isn't necessarily to perfectly recover the specific weight that past decision makers had implicitly assigned to different criteria, but rather to ensure that the model produces a similar set of decisions as humans. To return to one of our recurring examples, an educational institution might want to automate the process of grading essays, and it might attempt to do that by relying on machine learning to learn to mimic the grades teachers have assigned to similar work in the past.

This form of automation might help to address concerns with arbitrariness in human decision making by formalizing and fixing a decision-making scheme similar to what humans might have been employing in the past. In this respect, machine learning might be desirable because it can help to smooth out any inconsistencies in the human decisions from which it has induced some decision-making rule. For example, the essay grading model described above might reduce some of the variance observed in the grading of teachers whose subjective evaluations the model is learning to replicate. Automation can once again help to address concerns with arbitrariness understood as inconsistency, even when it is subjective judgments that are being automated.

A few decades ago, there was a popular approach to automation that relied on explicitly encoding the reasoning that humans relied on to make decisions.[18] This approach, called expert systems, failed for many reasons, including the fact that people aren't always able to explain their own reasoning.[19] Expert systems eventually gave way to the approach of simply asking people to label examples and having learning algorithms discover how to best predict the label that humans would assign. While this approach has proved powerful, it has its dangers.

First, it may give the veneer of objective assessment to decision-making schemes that simply automate the subjective judgment of humans. As a result, people may be more likely to view its decisions as less worthy of critical investigation. This is particularly worrisome because learning decision-making rules from the previous decisions made by humans runs the obvious risk of replicating and exaggerating any objectionable qualities of human decision making by learning from the bad examples set by humans. (In fact, many attempts to learn a rule to predict some seemingly objective target of interest — the form of automation that we'll discuss in the next section — are really just a version of replicating human judgment in disguise. If we can't obtain objective ground truth for the chosen target of prediction, there is no way to escape human judgment. As David Hand points out, humans will often need to exercise discretion in specifying and identifying what counts as an example of the target.[20])

Second, such decision-making schemes may be regarded as equivalent to those employed by humans and thus likely to operate in the same way, even though

the model might reach its decisions differently and produce quite different error patterns.[21] Even when the model is able to predict the decisions that humans would make given any particular input with a high degree of accuracy, there is no guarantee that the model will have inherited all of the nuance and considerations that go into human decision-making. Worse, models might also learn to rely on criteria in ways that humans would find worrisome or objectionable, even if doing so still produces a similar set of decisions as humans would make.[22] For example, a model that automates essay grading by assigning higher scores to papers that employ sophisticated vocabulary may do a reasonably good job replicating the judgments of human graders (likely because higher quality writing tends to rely on more sophisticated vocabulary), but checking for the presense of certain words is unlikely to be a reliable substitute for assessing an essay for logical coherence and factual correctness.[23]

In short, the use of machine learning to automate decisions previously performed by humans can be problematic because it can end up being both too much like human decision makers and too different from them.

Deriving Decision-Making Rules by Learning to Predict a Target

The final form of automation is one in which decision makers rely on machine learning to learn a decision-making rule or policy from data. This form of automation, which we'll call predictive optimization, speaks directly to concerns with reasoned decision making. Note that neither of the first two forms of automation does so. Consistently executing a pre-existing policy via automation does not ensure that the policy itself is a reasoned one. Nor does relying on past human decisions to induce a decision-making rule guarantee that the basis for automated decision making will reflect reasoned judgments. In both cases, the decision making scheme will only be as reasoned as the formal policy or informal judgments whose execution is being automated.

In contrast, predictive optimization tries to provide a more rigorous foundation for decision making by only relying on criteria to the extent that they demonstrably predict the outcome or quality of interest. When employed in this manner, machine learning seems to ensure reasoned decisions because the criteria that have been incorporated into the decision making scheme — and their particular weighing — are dictated by how well they predict the target. And so long as the chosen target is a good proxy for decision makers' goals, relying on criteria that predict this target to make decisions would seem well reasoned because doing so will help to achieve decision makers' goals.

Unlike the first two forms of automation, predictive optimization is a radical departure from the traditional approach to decision making. In the traditional approach, a set of decision makers has some goal — even if this goal is amorphous and hard to specify — and would like to develop an explicit decision-making scheme to help realize their goal. They engage in discussion and deliberation to try to come to some agreement about the criteria that are relevant to the decision and the weight to assign to each criterion in the decision-making scheme. Relying

on intuition, prior evidence, and normative reasoning, decision makers will choose and combine features in ways that are thought to help realize their goals.

The statistical or machine learning approach works differently. First, the decision makers try to identify an explicit target for prediction which they view as synonymous with their goal — or a reasonable proxy for it. In a college admissions scenario, one goal might be scholastic achievement in college, and college GPA might be a proxy for it. Once this is settled, the decision makers use data to discover which criteria to use and how to weight them in order to best predict the target. While they might exercise discretion in choosing the criteria to use, the weighting of these criteria would be dictated entirely by the goal of maximizing the accuracy of the resulting prediction of the chosen target. In other words, the decision-making rule would, in large part, be learned from data, rather than set down according to decision makers' subjective intuitions, expectations, and normative commitments.

Table 1: Comparison of traditional decision making to predictive optimization

	Traditional approach	Predictive optimization approach
Example: college admissions	Holistic approach that takes into account achievements, character, special circumstances, and other factors	Train a model based on past students' data to predict applicants' GPA if admitted; admit highest scoring applicants
Goal and target	No explicit target; goal is implicit (and there are usually multiple goals)	Define an explicit target; assume it is a good proxy for the goal
Focus of deliberation	Debate is about how the criteria should affect the decision	Debate is largely about the choice of target
Effectiveness	May fail to produce rules that meet their putative objectives	Predictive accuracy can be quantified
Range of normative considerations	Easier to incorporate multiple normative principles such as need	Harder to incorporate multiple normative principles
Justification	Can be difficult to divine rule makers' reasons for choosing a certain decision making scheme	Reasons for the chosen decision making scheme are made explicit in choice of target

Each approach has pluses and minuses from a normative perspective. The traditional approach makes it possible to express multiple goals and normative values through the choice of criteria and the weight assigned to them.

In the machine learning approach, multiple goals and normative considerations

need to be packed into the choice of target. In college admissions, those goals and considerations might include — in addition to scholastic potential — athletic and leadership potential, the extent to which the applicant would contribute to campus life, whether the applicant brings unusual life experiences, their degree of need, and many others. The most common approach is to define a composite target variable that linearly combines multiple components, but this quickly becomes unwieldy and is rarely subject to robust debate. There is also some room to exercise normative judgment about the choice to include or exclude certain decision criteria, but is a far cry from deliberative policy-making.

On the other hand, if we believe that a target does, in fact, capture the full range of goals that decision makers have in mind, machine learning models might be able to serve these goals more effectively. For example, in a paper that compares the two approaches to policy making, Rebecca Johnson and Simone Zhang show that the traditional approach (i.e., manually crafting rules via a process of deliberation and debate) often fails to produce rules that meet their putative objectives.[24] In examining rules for allocating housing assistance, they find that housing authorities prioritize veterans above particularly rent-burdened households, despite the fact that supporting such households would seem to be more in line with the policy's most basic goal. Johnson and Zhang assert that while this prioritization might be the actual intent of the policymakers setting the rules, the reasons for this prioritization are rarely made explicit in the process of deliberation and are especially difficult to discern after the fact. Were these rules developed instead using machine learning, policymakers would need to agree on an explicit target of prediction, which would leave much less room for confusion about policymakers' intent. And it would ensure that the resulting rules are *only* designed to predict that target.[24] As Rediet Abebe, Solon Barocas, Jon Kleinberg, and colleagues have argued, "[t]he nature of computing is such that it requires explicit choices about inputs, objectives, constraints, and assumptions in a system"[25] — and this may be a good thing if it forces certain policy considerations and normative judgements into the open.

The machine learning approach nevertheless runs the serious risk of focusing narrowly on the accuracy of predictions. In other words, "good" decisions are those that accurately predict the target. But decision making might be "good" for other reasons: focusing on the right qualities or outcomes (in other words, the target is a good proxy for the goal), considering only relevant factors, considering the full set of relevant factors, incorporating other normative principles (e.g., need, desert, etc.), or allowing people to understand and potentially contest the policy. Even a decision making process that is not terribly accurate might be seen as good if it has some of these other properties.[26] In the next few sections, we'll explore how each of these concerns might apply to machine learning.

Mismatch between target and goal

Identifying a target of prediction that is a good match for the goals of the decision maker is rarely straightforward. Decision makers often don't have a pre-existing, clear, and discrete goal in mind.[27] When they do, the goal can be far more complex and multifaceted than one discrete and easily measurable outcome.[28] In fact, decision makers can have multiple conflicting goals, perhaps involving some trade-offs between them. For example, the decision-making schemes adopted by college admission officers often encode a range of different goals. They do not simply rank applicants by their predicted grade point average and then admit the top candidates to fill the physical capacity of the school. Aside from the obvious fact that this would favor candidates who take "easy" classes, admissions officers aim to recruit a student body with diverse interests and a capacity to contribute to the broader community.

Besides, there might be serious practical challenges in measuring the true outcome of interest, leaving decision makers to find alternatives that might serve as a reasonable proxy for it. In most cases, decision makers settle on a target of convenience — that is, on a target for which there is easily accessible data.[8,29] For example, arrest data (i.e., whether someone has been arrested) is often adopted as a proxy for crime data (i.e., whether someone has committed a crime), even though many crimes are never observed and thus never result in arrest and even though the police might be quite selective in choosing to arrest someone for an observed crime.[30] Without condoning the decision to adopt this target, we might still recognize the practical challenges that would encourage the police to rely on arrests. It is simply impossible to observe all crime and so decision makers might feel justified in settling on arrests as a substitute.

Even if decision makers had some way of obtaining information on crime, it is *still* not obvious how well this chosen target would match the underlying goals of the police. Accurately predicting the occurrence of future crimes is not the same thing as helping to reduce crime; in fact, accurate predictions of crime might simply cause the police to observe more crimes and generate more arrests rather than preventing those crimes from happening in the first place.[31] If the police's actual goal is to reduce crime and not simply to ensure that all crimes result in arrests, then even using crime as the target of prediction might not help the police to realize these goals. The police might be better off estimating the deterrent effect of police intervention, but this is a far more complicated task than making predictions on the basis of observational data; answering these questions requires experimentation. (Of course, even this formulation of the problem should be subject to further critical analysis because it fails to consider the many other kinds of interventions that might help to reduce crime beyond improving the deterrent effect of police presence.) Yet even when there are good reasons to favor a more nuanced approach along these lines, decision makers may favor imperfect simplifications of the problem because they are less costly or more tractable.[13,8]

Finally, decision makers and decision subjects might have very different ideas about what would constitute the right target of prediction. Much of the discussion

in this chapter has so far been premised on the idea that decision makers' goals are widely perceived as desirable in the first place, and thus defensible. But there are many times when the normative issue is not with the way decisions are being made, but with the goal of the decision-making process itself.[29] In some cases, we may disagree with the goals of any given decision maker because we don't think that they are what is in the best interest of the decision makers themselves. More often, we might disagree with these goals because they are at odds with the interests of other people who will be negatively impacted by decision makers' pursuit of these goals. As Oscar Gandy has argued, "certain kind[s] of bias are inherent in the selection of the goals or objective functions that automated systems will [be] designed to support".[32]

To appreciate how this is different from a target-goal mismatch, consider a well-known study by Ziad Obermeyer, Brian Powers, Christine Vogeli, et al. on bias in an algorithm employed by a healthcare system to predict which patients would benefit most from a "high-risk care management" program.[33] They found that algorithm exhibited racial bias — specifically, that it underestimated the degree to which black patients' health would benefit from enrollment in the program. That's because the developers adopted healthcare costs as the target of prediction, on the apparent belief that it would serve as a reasonable proxy for healthcare needs. The common recounting of this story suggests that decision makers simply failed to recognize the fact that there are racial disparities in both care-seeking and the provision of healthcare that cause black patients of equivalently poor health to be less expensive than non-black patients. On this account, fixing the problem would only require adopting a target that better reflected the healthcare system's goals: maximizing the overall health benefits of the program. Yet it is entirely possible that the original target of prediction reflected the healthcare system's true goals, which might have been to simply reduce costs without any regard for whose health would benefit most from these interventions. If that were the case, then the choice of target was not simply a poor match for decision makers' goals; the goals themselves were problematic. We must be careful not to confuse cases where we object to the goals for cases where we object to the particular choice of target.

Failing to consider relevant information

Bureaucracies are often criticized for not being sufficiently individualized or particularized in their assessments, lumping people into needlessly coarse groups. Had decision makers only considered some additional detail, they would have realized that the person in question is actually unlike the rest of the people with whom they have been grouped.

Supervised machine learning is a form of inductive reasoning. It aims to draw general rules from a set of specific examples, identifying the features and feature values that reliably co-occur with an outcome of interest. As it turns out, the limitation of being insufficiently individualized is an unavoidable part of inductive reasoning.

Imagine a car insurance company that is trying to predict the likelihood that a person applying for an insurance policy will get into a costly accident. The insurer will try to answer this question by looking at the frequency of past accidents that involved other people similar to the applicant. This is inductive reasoning: the applicant is likely to exhibit similar behavior or experience similar outcomes as previous policyholders because the applicant possesses many other qualities in common with these policyholders. Perhaps the person is applying for insurance to cover their bright red sports car — a type of car that is involved in accidents far more frequently than other types of cars. Noting this historical pattern, the insurer might therefore conclude that there is a heightened chance that the applicant will need to make a claim against their policy — and only offer to insure the applicant at an elevated price. Having received the offer, the applicant, who is, in fact, a highly skilled driver with an excellent command of the car, might balk at the price, objecting to the idea that they present a risk anything like the other policyholders with the same car.

What has happened here? The insurer has made its prediction on the basis of rather coarse details (in this case, on the basis of only the model and color of the car), treating the rate at which accidents happen among previous policyholders with such a car as a reliable indicator of the probability of the applicant having an accident of their own. Frederick Schauer refers to this as the problem of "statistically sound but nonuniversal generalizations": when an individual fulfills all the criteria for inclusion in a particular group, but fails to possess the quality that these criteria are expected to predict.[34]

Situations of this sort can give rise to claims of stereotyping or profiling and to demands that decision makers assess people as individuals, not merely as members of a group. Yet, as Schauer has explained, it can be difficult to specify what it means to treat someone as an individual or to make individualized decisions. It is unclear, for example, how an insurer could make predictions about an individual's likelihood of getting into a car accident without comparing the applicant to other people that resemble them. At issue in these cases is not the failure to treat someone as an individual, but the failure to take additional relevant criteria into account that would distinguish a person from the other people with whom they would otherwise be lumped in with.[34] If the insurer had access to additional details (in particular, details about the applicant's and past policyholders' driving skills), the insurer might have made a more discerning judgment about the applicant. This is exactly what is going on when insurers agree to offer lower prices to applicants who voluntarily install black boxes in their cars and who demonstrate themselves to be careful drivers. It is easy to misinterpret this trend as a move toward individualized assessment, as if insurers are judging each individual person on their unique merits as a driver. The correct interpretation requires that we recognize that insurers are only able to make use of the data from a specific driver's black box by comparing it to the data from the black boxes of other drivers whose driving records are being used to make a prediction about the driver in question. Even if we accept that decisions can never be fully individualized, we might still expect that decision makers take into account the full range of relevant information at their potential

disposal. To carry forward the example above, we might say that the car insurance company had an obligation to consider the applicants' driving skills, not just the model and color of their car, even if doing so still meant that they were being assess according to how often other people with similar driving skills and similar cars have gotten into accidents in the past.

But how far should this expectation extend? What obligations do decision makers have to seek out every last bit of conceivable information that might enable more accurate predictions? Well, at some point, additional information ceases to be helpful because there isn't enough training data. For example, people who live near a specific intersection may be more likely to get into accidents because the intersection is poorly designed and thus dangerous. But the insurer can only learn this if it has enough data from enough people who live near this intersection.

There is also a very practical reason why we might not hold decision makers to a standard in which they are required to consider all information that might be conceivably relevant. Collecting and considering all of this information can be expensive, intrusive, and impractical. In fact, the cost of doing so could easily outweigh the perceived benefits that come from more granular decision making — not just to the decision maker but to the subjects of the decisions as well. While black boxes can help to achieve far more granularity in insurance pricing, they are also controversial because they are quite intrusive and pose a threat to drivers' privacy. For reasons along these lines, Schauer and others have suggested that decision makers are justified in making decisions on the basis of a limited set of information, even when additional relevant information might exist, if the cost of obtaining that information swamps out its benefits.[34,35]

There are three things to note about these arguments. First, these are not arguments about automated decision-making specifically; they are general statements about any form of decision making, whether automated or not. Yet, as we discussed earlier in the chapter, automated decision making often limits the opportunity to introduce additional relevant information into the decision-making process. The cost-savings that might be achieved by automating certain decisions (often by way of replacing human workers with software) comes at the cost of depriving people the chance to highlight relevant information that has no place in the automated process. Given that people might be both very willing and perfectly able to volunteer this information (i.e., able to do so at little cost), automated decision-making that simply denies people the opportunity to do so might fail the cost-benefit analysis. Second, the cost-benefit analysis that undergirds Schauer and others' arguments does not take into account any distributional considerations, like which groups might be enjoying more of the benefits or experiencing more of the costs. In Chapter 4, we'll return to this question, asking whether decision makers are justified in subjecting certain groups to less granular and thus less accurate decisions simply because there is less information about them. Finally, these arguments don't grapple with the fact that decision makers and decision subjects might arrive at quite different conclusions when performing a cost-benefit analysis if they are performing this analysis from their own perspectives. A decision-maker might find that the costs of collecting more information does not generate a sufficiently

large corresponding benefit *for them as the decision maker*, despite the fact that certain decision subjects would surely benefit from such an investment. It is not obvious why the cost-benefit analysis of decision makers alone should be allowed to determine the level of granularity that is acceptable. One possible explanation might be that increasing the costs of making decisions (by, for example, seeking out and taking more information into account) will encourage decision makers to simply pass these costs onto decision subjects. For instance, if developing a much more detailed assessment of applicants for car insurance increases the operating costs of the insurer, the insurer is likely to charge applicants a higher price to offset these additional costs. From this perspective, the costs to the decision maker are really just costs to decision subjects. Of course, this perspective doesn't contemplate the possibility of the insurer simply assuming these costs and accepting less profit.

The limits of induction

Beyond cost considerations, there are other limits to inductive reasoning. Suppose the coach of a track team assesses potential members of the team according to the color of their sneakers rather than the speed with which they can run. Imagine that just by coincidence, slower runners in the pool happen to prefer red sneakers and faster runners happen to prefer blue sneakers — but that no such relationship obtains in other pools of runners. Thus, any lessons the coach might draw from these particular runners about the relationship between sneaker colors and speed would be unreliable when applied to other runners. This is the problem of overfitting.[36] It is a form of arbitrary decision making because the predictive validity that serves as its justification is an illusion.

Overfitting is a well-understood problem in machine learning and there are many ways to counteract it. Since the spurious relationship occurs due to coincidence, the bigger the sample, the less likely it is to occur. Further, one can penalize models that are overly complicated to make it less likely that they pick up on chance patterns in the data. And most importantly, it is standard practice to separate the examples that are used to train and test machine learning models. This allows a realistic assessment of how well the relationships observed in the training data carry over to unseen examples. For these reasons, unless dealing with small sample sizes, overfitting is generally not a serious problem in practice.

But variants of the overfitting problem can be much more severe and thorny. It is common practice in machine learning to take one existing dataset — in which all the data has been gathered in a similar way — and simply split this dataset into training and test sets. The small differences between these sets will help to avoid overfitting and may give some sense for performance on unseen data. But these splits are still much more similar to each other than the future population to which the model might be applied.[37,38] This is the problem of "distribution shifts," of which there are many different kinds. They are common in practice and they present a foundational problem for the machine learning paradigm.

Returning to our earlier example, imagine that runners are only able to buy

sneakers from one supplier and that the supplier only sells one type of sneaker, but varies the color of the sneaker by size (all sizes below 8 are red, while all sizes 8 and above are blue). Further, assume that runners with larger feet are faster than those with smaller feet and that there is a large step change in runners' speed once their foot size exceeds 7. Under these circumstances, selecting runners according to the color of their sneakers will reliably result in a team composed of faster runners, but it will do so for reasons that we still might find foolish or even objectionable. Why? The relationship between the color of a runner's sneakers and running speed is obviously spurious in the sense that we know that the color of a runner's sneakers has no causal effect on speed. But is this relationship truly spurious? It is not just an artifact of the particular set of examples from which a general rule has been induced; it's a stable relationship in the real world. So long as there remains only one supplier and the supplier only offers different colors in these specific sizes, sneaker color will reliably distinguish faster runners from slower runners. So what's the problem with making decisions on this basis? Well, we might not always have a way to determine whether we are operating under the conditions described. Generalizing from specific examples always admits the possibility of drawing lessons that do not apply to the situation that decision makers will confront in the future.

One response to these concerns is to assert that there is a normative obligation that decision criteria bear a causal relationship to the outcome that they are being used to predict. The problem with using sneaker color as a criterion is obvious to us because we can recognize the complete absence of any plausible causal influence on running speed. When machine learning is used, the resulting models, unconcerned with causality, may seize upon unstable correlations.[39] This gives rise to demands that no one should be subject to decision-making schemes that are based on findings that lack scientific merit — that is, on findings that are spurious and thus invalid. They likely account for concerns of scholars like Frank Pasquale, who talks about cases where machine learning is "facially invalid",[40] and Pauline Kim and Erika Hanson, who have argued that "because data mining uncovers statistical relationships that may not be causal, relying on those correlations to make predictions about future cases may result in arbitrary treatment of individuals".[41] Asserting that decision-making schemes should only be based on criteria that have a causal relationship to the outcome of interest are likely perceived as a way to avoid these situations — that is, as a way to ensure that the basis for decision making is well reasoned, not arbitrary.

A right to accurate predictions?

In the previous two sections, we discussed several reasons why predictions using inductive reasoning may be inaccurate, including failing to consider relevant information and distribution shift. But even if we set aside those reasons — assume that the decision maker considers all available information, there is no distribution shift, etc. — there might be insurmountable limits to the accuracy of predicting

future outcomes. These limits might persist whether or not inductive reasoning is employed.[42] For example, at least some cases of recidivism are due to spur-of-the-moment crimes committed when opportunities fortuitously presented themselves, and these might not be predictable in advance. (We'll review some of the empirical evidence of limits to prediction in later chapters.)

What are the implications of these limits to prediction? From the decision maker's perspective, even a small increase in predictive accuracy compared to a baseline (human judgment or rule-based policy) can be valuable. Consider a child protection agency employing a predictive screening tool to determine which children are at risk of child abuse. Increased accuracy may mean fewer children placed in foster care. It might also result in substantial cost savings, with fewer caseworkers required to make visits to homes.

A typical model deployed in practice may have an accuracy (more precisely, AUC) of between 0.7 and .8.[43] That's better than a coin toss but still results in a substantial number of false positives and false negatives. A claim that the system makes the most accurate decision possible at the time of screening is cold comfort to families where children are separated from their parents due to the model's prediction of future abuse, or cases of abuse that the model predicted to be low risk. If the model's outputs were random, we would clearly consider it arbitrary and illegitimate (and even cruel). But what is the accuracy threshold for legitimacy? In other words, how high must accuracy be in order to be able to justify the use of a predictive system at all?[44]

Low accuracy becomes even more problematic when we consider that it is measured with respect to a prediction target that typically requires sacrificing some of the multifaceted goals that decision makers might have. For example, a child welfare risk prediction model might not be able to reason about the differential effects that an intervention such as foster care might have on different children and families. How much of an increase in predictive accuracy is needed to justify the mismatch between the actual goals of the system and those realized by the model?

Obviously, these questions don't have easy answers, but they represent important and underappreciated threats to the legitimacy of predictive decision making.

Agency, recourse, and culpability

Let's now consider a very different concern: could criteria that exhibit statistical relevance and enable accurate predictions still be normatively inappropriate as the basis for decision making?

Perhaps the criterion in question is an immutable characteristic. Perhaps it is a mutable characteristic, but not one that the specific person in question has any capacity to change. Or perhaps the characteristic has been affected by the actions of others, and is not the result of the person's own actions. Each of these reasons, in slightly different ways, all concern the degree of control that a person is understood to have over the characteristic in question — and each provides some normative justification for either ignoring or discounting the characteristic even

when it might be demonstrably predictive of the outcome of interest. Let's dig into each of these concerns further.

Decisions based on immutable characteristics can be cause for concern because they threaten people's agency. By definition, there is nothing anyone can do to change immutable characteristics (e.g., one's country of birth). By extension, there is nothing anyone can do to change decisions made on the basis of immutable characteristics. Under these circumstances, people are condemned to their fates and are no longer an agent of their own lives. There is something disquieting about the idea of depriving people of the capacity to make changes that would result in a different outcome from the decision-making process, especially when these decisions might significantly affect a person's life chances and life course. This might be viewed as especially problematic when there seem to be alternative ways for a decision maker to render effective judgment about a person without relying on immutable characteristics. In this view, if it is possible to develop decision-making schemes that are equally accurate, but still leave room for decision subjects to adapt their behavior so as to improve their chances of a favorable decision, then decision makers have an moral obligation to adopt such a scheme out of respect for people's agency.

Recourse is a related but more general idea about the degree to which people have the capacity to make changes that result in different decisions.[45] While there is nothing anyone might do to change an immutable characteristic, people might be more or less capable of changing those characteristics that are, in principle, mutable.[46,47] Some people might need to expend far more resources than others to obtain the outcome that they want from the decision-making process. Choosing certain criteria to serve as the basis for decision making is also a choice about the kinds of actions that will be available for people to undertake in seeking a different decision. And people in different circumstances will have different abilities to successfully do so. In some cases, people may never have sufficient resources to achieve this — bringing us back to the same situation discussed in the previous paragraph. For example, one applicant for credit might be well positioned to move to a new neighborhood so as to make herself a more attractive candidate for a new loan, assuming that the decision making scheme uses location as an important criterion. But another applicant might not be able to do so, for financial, cultural, or many other reasons.

Research on recourse in machine learning has largely focused on ensuring that people receive *explanations* of ways to achieve a different decision from a model that people can actually execute in reality.[48] Given that there are many possible ways to explain the decisions of a machine learning model, the goal of this work is to ensure that the proffered explanations direct people to take viable actions rather than suggesting that the only way to get the desired outcome is to do something beyond their capacity. Even when developing a decision-making scheme that only relies on mutable characteristics, decision makers can do more to preserve recourse by adapting their explanation of a model's decisions to focus on those actions that are easiest for people to change. On this account, the better able people are to make changes that give them the desired outcome, the better the decision-making

scheme and the better the explanation.

Finally, as mentioned earlier in this section, we might view certain decision-making schemes as unfair if they hold people accountable for characteristics outside their control. Basic ideas about moral culpability almost always rest on some understanding of the actions that brought about the outcomes of concern. For example, we might be upset with a person who has bumped into us and caused us to drop and break some precious item. Upon discovering that they have been pushed by somebody else, we are likely to hold them blameless and redirect our disapprobation to the person who pushed them. This same reasoning often carries over to the way that we think about the fairness of relying on certain criteria when making decisions that allocate desirable resources and opportunities. Unless we know *why* certain outcomes come to pass, we cannot judge whether decision makers are normatively justified in relying on criteria that accurately predict if that outcome will come to pass. We need to understand the cause of the outcome of interest so that we might reflect on whether the subject of the decision bears moral responsibility for the outcome, given its cause.

For example, as Barbara Kiviat has explored, laws in many U.S. states limit the degree to which car insurance providers can take into account "extraordinary life circumstances" when making underwriting or pricing decisions, including such events as the death of a spouse, child, or parent.[49] These laws forbid insurers from considering a range of factors over which people cannot exercise any control — like a death in the family — but which may nevertheless contribute to someone experiencing financial hardship and thus to increasing the likelihood of making a claim against their car insurance policy in the event of even a minor accident. These prohibitions reflect an underlying belief that people should not be subject to adverse decisions if they were not responsible for whatever it is that makes them appear less deserving of more favorable treatment. Or to put it another way: people should only be judged on the basis of factors for which they are morally culpable. Fully implementing this principle is impractical, since most attributes that the decision maker might use, say income, are partly but not fully the result of the individual's choices. However, attributes like a death in the family seem to fall fairly clearly on one side of the line.

Of course, there is a flip side to all of this. If people can easily change the features that are used to make decisions about them, they might "game" the decision-making process. By gaming we mean changing the value of features in order to change the decision without changing the expected outcome that the features are meant to predict.[50] 'Teaching to the test' is a familiar scenario that is an example of gaming. Here, the test score is a feature that predicts future performance (say, at a job). Assume that the test, in fact, has predictive value, because people who do well at the test tend to have mastered some underlying body of knowledge, and such mastery improves their job performance. Teaching to the test refers to methods of preparation that increase the test score without correspondingly increasing the underlying ability that the score is meant to reflect. For example, teachers might help students prepare for the test by exploiting the fact that the test assesses very specific knowledge or competencies — not the full

range of knowledge or competencies that the test purports to measure — and focus preparation on only those parts that will be assessed.[51] Jane Bambauer and Tal Zarsky give many examples of gaming decision making systems.[52]

Gaming is a common problem because most models do not discover the causal mechanism that accounts for the outcome. Thus, preventing gaming requires causal modeling.[50] Furthermore, a gameable scheme becomes less effective over time and may undermine the goals of the decision maker and the proper allocation of the resource. In fact, gaming can be a problem even when decision subjects are not acting adversarially. Job seekers may expend considerable effort and money to obtain meaningless credentials that they are told matter in their industry, only to find that while this helps them land a job, it does not make them any better prepared to actually perform it.[53] Under such circumstances, strategic behavior may represent wasteful investment of effort on the part of well-intentioned actors.

Concluding thoughts

In this chapter, we teased apart three forms of automation. We discussed how each of these responds to concerns about arbitrary decision making in some ways, while at the same time opening up new concerns about legitimacy. We then delved deep into the third type of automation, predictive optimization, which is what we'll be concerned with in most of this book.

To be clear, we make no blanket claims about the legitimacy of automated decision making or predictive optimization. In applications that aren't consequential to people's life chances, questions of legitimacy are less salient. For example, in credit card fraud detection, statistical models are used to find patterns in transaction data, such as a sudden change in location, that might indicate fraud resulting from stolen credit card information. The stakes to individuals tend to be quite low. For example, in the United States, individual liability is capped at $50 provided certain conditions are met. Thus, while errors are costly, the cost is primarily borne by the decision maker (in this example, the bank). So banks tend to deploy such models based on cost considerations without worrying about (for instance) providing a way for customers to contest the model.

In consequential applications, however, to establish legitimacy, decision makers must be able to affirmatively justify their scheme along the dimensions we've laid out: explain how the target relates to goals that all stakeholders can agree on; validate the accuracy of the deployed system; allow methods for recourse, and so forth. In many cases, it is possible to put procedural protections around automated systems to achieve this justification, yet decision makers are loath to do so because it undercuts the cost savings that automation is meant to achieve.

Bibliography

[1] Konger, Kate and Chen, Brian X.. 2022. "A change by Apple is tormenting Internet companies, especially Meta". The New York Times https://www.nytimes.com/2022/02/03/technology/apple-privacy-changes-meta.html.

[2] Richardson, Willliam Jamal. 2017. "Against black inclusion in facial recognition".

[3] Powles, Julia and Nissenbaum, Helen. 2018. "The seductive diversion of 'solving' bias in artificial intelligence". Medium.

[4] Weber, Max. 2019. *Economy and Society*. Harvard University Press, Cambridge, MA.

[5] Strandburg, Katherine J.. 2019. "Rulemaking and inscrutable automated decision tools". *Columbia Law Review*, 119(7):1851–1886.

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

[7] Kroll, Joshua A., Huey, Joanna, Barocas, Solon, Felten, Edward W., Reidenberg, Joel R., Robinson, David G., and Yu, Harlan. 2017. "Accountable algorithms". *University of Pennsylvania Law Review*, 165(3):633–705.

[8] Citron, Danielle Keats. 2008. "Technological Due Process". *Washington University Law Review*, 85(6):1249–1313.

[9] Christie, James. 2020. "The post office horizon it scandal and the presumption of the dependability of computer evidence". *Digital Evidence & Elec. Signature L. Rev.*, 17:49.

[10] Kaplow, Louis. 1992. "Rules versus Standards: An Economic Analysis". *Duke Law Journal*, 42(3):557–629.

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

[12] Colin Lecher. 2018. "What happens when an algorithm cuts your health care". https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[13] Clarke, Roger. 1988. "Information technology and dataveillance". *Communications of the ACM*, 31(5):498–512.

[14] Kaminski, Margot E. and Urban, Jennifer M.. 2021. "The right to contest AI". *Columbia Law Review*, 121(7):1957–2048.

[15] Gilman, Michele. 2020. "Poverty lawgorithms". Technical report, Data & Society, New York, NY.

[16] Nissenbaum, Helen. 1996. "Accountability in a computerized society". *Science and Engineering Ethics*, 2(1):25–42.

[17] Binns, Reuben, Kleek, Max Van, Veale, Michael, Lyngs, Ulrik, Zhao, Jun, and Shadbolt, Nigel. 2018. "'It's Reducing a Human Being to a Percentage': Perceptions of justice in algorithmic decisions". *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[18] Collins, Harry. 1991. *Artificial Experts: Social Knowledge and Intelligent Machines*. MIT Press, Cambridge, MA.

[19] Forsythe, Diana E.. 2000. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford University Press, Stanford, CA.

[20] Hand, David J.. 2006. "Classifier Technology and the Illusion of Progress". *Statistical Science*, 21(1):1–14.

[21] Burrell, Jenna. 2016. "How the machine 'thinks': Understanding opacity in machine learning algorithms". *Big Data & Society*, 3(1).

[22] Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

[23] Perelman, Les. 2012. "Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (aes)". *International advances in writing research: Cultures, places, measures*, pages 121–131.

[24] Johnson, Rebecca Ann and Zhang, Simone. 2022. "What is the bureaucratic counterfactual? categorical versus algorithmic prioritization in U.S. social policy". pages 1671–1682.

[25] Abebe, Rediet, Barocas, Solon, Kleinberg, Jon, Levy, Karen, Raghavan, Manish, and Robinson, David G. 2020. "Roles for computing in social change". pages 252–260.

[26] Tyler, Tom R. 1988. "What is Procedural Justice?: Criteria used by Citizens to Assess the Fairness of Legal Procedures". *Law & Society Review*, 22(1):103–135.

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

[28] Hand, David J.. 1994. "Deconstructing Statistical Questions". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157(3):317–338.

[29] Richardson, Rashida. 2022. "Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities". *Berkeley Technology Law Journal*, 36(3):1051–1090.

[30] Lum, Kristian and Isaac, William. 2016. "To predict and serve?" *Significance*, 13(5):14–19.

[31] Harcourt, Bernard E. 2008. *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

[32] Gandy, Oscar H.. 2010. "Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems". *Ethics and Information Technology*, 12(1):29–42.

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

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

[35] Lippert-Rasmussen, Kasper. 2011. ""We are all Different": Statistical Discrimination and the Right to be Treated as an Individual". *The Journal of Ethics*, 15(1-2):47–59.

[36] Mitchell, Tom M.. 1980. "The need for biases in learning generalizations". Technical report, Department of Computer Science, Laboratory for Computer Science Research, Rutgers University, New Brunswick, NJ.

[37] Oreskes, Naomi, Shrader-Frechette, Kristin, and Belitz, Kenneth. 1994. "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences". *Science*, 263(5147):641–646.

[38] Malik, Momin M. 2020. "A hierarchy of limitations in machine learning". *arXiv preprint arXiv:2002.05193*.

[39] Mayer-Schönberger, Viktor and Cukier, Kenneth. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Harper Business, New York, NY.

[40] Pasquale, Frank. 2018. "When machine learning is facially invalid". *Communications of the ACM*, 61(9):25–27.

[41] Kim, Pauline T. and Hanson, Erika. 2016. "People analytics and the regulation of information under the Fair Credit Reporting Act". *Saint Louis University Law Journal*, 61(1):17–34.

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

[43] Chouldechova, Alexandra, Putnam-Hornstein, Emily, Benavides-Prado, Diana, Fialko, Oleksandr, and Vaithianathan, Rhema. 2017. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions". In *Proceedings of Machine Learning Research*, volume 81, pages 1–15.

[44] Huq, Aziz Z.. 2020. "A Right to a Human Decision". *Virginia Law Review*, 106(3):611–688.

[45] Ustun, Berk, Spangher, Alexander, and Liu, Yang. 2019. "Actionable Recourse in Linear Classification". In *Conference on Fairness, Accountability, and Transparency*, pages 10–19.

[46] Milli, Smitha, Miller, John, Dragan, Anca D., and Hardt, Moritz. 2019. "The Social Cost of Strategic Classification". In *Conference on Fairness, Accountability, and Transparency*, pages 230–239.

[47] Hu, Lily, Immorlica, Nicole, and Vaughan, Jennifer Wortman. 2019. "The Disparate Effects of Strategic Manipulation". In *Conference on Fairness, Accountability, and Transparency*, pages 259–268.

[48] Karimi, Amir-Hossein, Barthe, Gilles, Schölkopf, Bernhard, and Valera, Isabel. 2022. "A survey of algorithmic recourse: contrastive explanations and consequential recommendations". *ACM Computing Surveys*.

[49] Kiviat, Barbara. 2019. "The moral limits of predictive practices: The case of credit-based insurance scores". *American Sociological Review*, 84(6):1134–1158.

[50] Miller, John, Milli, Smitha, and Hardt, Moritz. 2020. "Strategic classification is causal modeling in disguise". In *Proceedings of the 37th International Conference on Machine Learning*, pages 6917–6926.

[51] O'Neil, Cathy. 2017. *Weapons of Math Destruction*. Penguin Random House, New York, NY.

[52] Bambauer, Jane and Zarsky, Tal. 2018. "The algorithm game". *Notre Dame L. Rev.*, 94:1.

[53] Chen, Irene Y., III, Hal Daumé, and Barocas, Solon. 2021. "The many roles that causal reasoning plays in reasoning about fairness in machine learning". In *NeurIPS Workshop on Algorithmic Fairness through the lens of Causality and Robustness*.