Algorithms as discrimination detectors

Jon Kleinberg[a,1], Jens Ludwig[b], Sendhil Mullainathan[c], and Cass R. Sunstein[d]

[a]Department of Computer Science, Cornell University, Ithaca, NY 14853; [b]Harris School of Public Policy, University of Chicago, Chicago, IL 60637; [c]Booth School of Business, University of Chicago, Chicago, IL 60637; and [d]Harvard Law School, Harvard University, Cambridge, MA 02138

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved June 9, 2020 (received for review September 18, 2019)

Preventing discrimination requires that we have means of detecting it, and this can be enormously difficult when human beings are making the underlying decisions. As applied today, algorithms can increase the risk of discrimination. But as we argue here, algorithms by their nature require a far greater level of specificity than is usually possible with human decision making, and this specificity makes it possible to probe aspects of the decision in additional ways. With the right changes to legal and regulatory systems, algorithms can thus potentially make it easier to detect—and hence to help prevent—discrimination.

machine learning | algorithms | discrimination

If we cannot detect discrimination, we cannot prevent it. Yet detecting it can be enormously difficult in practice. Statistical studies show that discrimination is widespread. But statistical evidence, such as lower call-back rates for resumes from African-American job applicants (1, 2), is aggregate by its very nature. It typically does not show that specific firms or individuals have discriminated against particular people. Without evidence to that effect, legal challenges to discrimination become quite difficult under existing law. And when such legal challenges are made, concrete evidence is hard to come by. Courts often struggle to determine whether a given outcome is a product of discriminatory intentions on the part of specific actors or attributable to some more benign explanation.

The struggle is intensified by the frequent opacity of human decision making and the difficulty of understanding even what we ourselves do and why. The best indicator of the seriousness of this struggle to detect discrimination comes from the continued prevalence of discrimination in statistical studies. We see the litter clearly, but catching the litterers is hard.

Algorithms have now been added to the mix and much attention has focused on the risk that they too might discriminate. No one should doubt that the risk is real. Algorithms are built by humans. They are trained on data generated by humans. Humans discriminate, and so the algorithms they construct can discriminate as well. A growing body of research has documented some of the pernicious ways in which this process plays out in practice (e.g., refs. 3–5). But the net effect of algorithms on society depends ultimately not just on whether they discriminate, but also on how they affect this core problem of detecting (and hence preventing) discrimination when it happens.

This essay, which summarizes and builds on our arguments in Kleinberg et al. (6), makes two main points. First, the existing legal, regulatory, and related systems for detecting discrimination were originally built for a world of human decision makers, unaided by algorithms. Without changes to these systems, the introduction of algorithms will not help with the challenge of detecting discrimination and could potentially make the whole problem worse. Our second point is more optimistic: Algorithms by their nature require a far greater level of specificity than is usually involved with human decision making, which in some sense is the ultimate "black box." With the right legal and regulatory systems in place, algorithms can serve as something akin to a Geiger counter that makes it easier to detect—and hence prevent—discrimination.

We emphasize that this is an aspiration, not a prediction. A great deal depends on putting the right legal and regulatory systems in place. Aspirationally, such systems can help not only in detecting discrimination as it is now understood in law, but also in clarifying the normative questions raised by debates over that contested concept. Is it discriminatory to adopt a neutral practice (such as a height requirement for a certain job) that has a disparate impact on women? What can be done to test whether practices of that kind can be justified? And what if a private or public institution uses a factor (such as credit ratings or criminal records) that may be bound up with past discrimination? The use of algorithms cannot answer such questions, but it can help produce another level of clarity about the stakes and about potential tradeoffs.

For all these reasons, formulating appropriate legal and regulatory structures for algorithmic decision making will be crucial: Without these structures we risk a great deal of discriminatory harm by algorithms; and with them we have the potential to forestall discrimination done by humans.

The Challenge of Detecting Human Discrimination Directly from Humans

To see why it can be so difficult to detect bias in an entirely human-driven decision system, it is useful to consider a concrete, and stylized, example. Suppose that a technology (tech) firm is accused of gender discrimination in hiring, in one of the 73 countries that prohibit such discrimination.* There is no question the firm has hired fewer women than men in the past, but the question at the heart of the litigation is why.

The plaintiff claims that the firm intentionally discriminated against women. The firm responds that there are differences in the applicant pool: Fewer women major in Science, Technology, Engineering, and Mathematics (STEM) fields.† The firm acknowledges that these disparities in STEM preparation could be due to gender bias somewhere—for example by teachers or parents—but argue that wherever it is, it is upstream from them. The firm adds that it did not intend to discriminate in any way. (Under existing law, at least in the United States, that could be an adequate defense against a charge of "disparate treatment," a principal form of discrimination.)

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "The Science of Deep Learning," held March 13–14, 2019, at the National Academy of Sciences in Washington, DC. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019 colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/science-of-deep-learning.

Author contributions: J.K., J.L., S.M., and C.R.S. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: kleinberg@cornell.edu.

First published July 28, 2020.

*https://wbl.worldbank.org/.

†Stoet and Geary (7) find that in a majority of countries girls score as well as boys in science, but in almost every country women underenroll in STEM fields relative to the share capable of doing college-level STEM work.

The plaintiff believes that the firm is lying and starts off by doing the obvious: asking the heads of human resources (HR) to testify and directly asking them whether they meant to discriminate against women. This initially seems like a promising approach. The heads will be testifying in court and so will face some penalty for lying. And we all tend to believe that humans can meaningfully answer questions about what they did and why. In this case the HR managers claim they did not discriminate. The plaintiff's lawyer asks, "Would my client have been hired had she been a man?" Again, they answer no.

But what can we actually conclude from this? One possibility is that they are indeed lying and know that they are doing so. But how can we know? In addition, a key finding from the last several decades of research in psychology is that the HR managers might themselves not really understand how they decided or why (see for example ref. 8). Dual-processing theories in psychology suggest that two sets of cognitive operations make up human cognition: 1) deliberate, conscious, effortful thought (often called "system II thinking") and 2) rapid, automatic, noneffortful responses that we may not even be aware of while they are taking place ("system I thinking").[‡] Because of the differences in effort required, we tend to use the automatic systems to save mental energy (12). Thus, for example, we might automatically think "two" when we hear "one plus one." Importantly, what our automatic system I is doing, and why, is frequently not accessible to our conscious system II thinking.

A person's automatic processing can potentially even conflict with that person's conscious thoughts, including when decisions involving the risk of discrimination take place.[§] People have a powerful tendency to categorize others and to prefer "in-groups" to "out-groups"; our automatic systems pay close attention to the age, race, and sex of others (16). The result can be implicit biases about which we are often unaware ourselves. In many ways, human cognition forms the ultimate black box, even to the person engaging in the cognitive activity.

Returning to our stylized example, the plaintiff's lawyer shifts gears and interrogates data for statistical evidence of discrimination. Are less productive men sometimes hired over more productive women? But it turns out that this firm's hiring process (like many screening decisions) does not involve any sort of formula or guidelines. The HR team is asked to make subjective, holistic assessments of applicants. The HR managers cannot define what they mean when they say they are looking for "productive workers."

The discussion shifts to qualifications instead: How do candidates with the same qualifications but different genders get treated? Answering this question stumbles on the follow-up questions of which qualifications matter. The plaintiff tries to compare candidates with the same qualifications on every possible dimension, but this turns out to be a long list. Even in large companies, statistical proof of intentional discrimination can be challenging to produce. And in countries like the United States and many others, small businesses account for the majority of hiring. So as it turns out, with the number of people the firm hires there are no two perfectly comparable applicants.

Challenges in using statistical evidence to show intentional discrimination, small sample sizes, unclear objectives, and the general opacity of human cognition combine to create a fog of ambiguity, which prevents us from stopping a behavior that we know to be widespread yet for which in any one instance there may well be plausible alternative explanations.

Defining Our Scope

To understand how algorithms might affect the current state of affairs, we now need to be more specific about what we mean. The word "algorithms" encompasses a wide range of tools from optimization to search, which in turn influence a variety of decisions.

The hiring example above illustrates our focus: a type of decision that is made billions of times per year all across the world in domains like credit lending, school admissions, or diagnosis in health care. In the hiring case, one or more people need to be selected from a larger candidate pool to optimize some outcome. The challenge is that the outcome is not known for the specific candidates at the time the decision is made. Which applicants will turn out to be the most productive workers if hired? In the credit case, which borrowers will default on their loans if given credit? So the missing outcomes must be guessed at—that is, predicted—based on what we know about the candidates at the time of the decision. Such screening decisions do not capture all of the potential uses of algorithms for decision making, but are among the most dominant uses of algorithms.[¶]

In this context, we focus on algorithms that build prediction functions from training data; the prediction function produced, in turn, takes "inputs" (like the characteristics of a college applicant) and predicts some outcome (like college grade point average). Such algorithms are generally developed using the methods of machine learning; beyond this level of generality there is nothing specific about our argument that hinges on any particular formalism within machine learning—for example, whether we are working with, say, a neural network or some other machine-learning technique. Regardless of the complexity of the function class that the learning procedure allows, all standard techniques share the property that the algorithm builder will need to specify an outcome to be predicted, what candidate predictors will be made available to the learning procedure, and a dataset.

We intentionally discuss these issues in the abstract, not grounded in any specific legal or regulatory system. The reason for this level of abstraction is pragmatic. The question of how to adapt to the growing use of algorithms is a global challenge, not specific to any particular country, and laws vary dramatically across countries. Moreover, laws change over time as technologies change. The US Constitution does not have much to say about wiretapping; the need to consider how to apply constitutional protections to wiretapping came about only once there were telephone wires that might be tapped.

Notwithstanding this point, we are keenly aware that the idea of "discrimination" is contested and ambiguous. Different legal systems understand the idea in different ways. In many legal systems, it is discriminatory to choose to treat people differently and, worse, because of some protected characteristic (race, religion, sex, age). We have focused thus far on intentional discrimination of this kind, which is the canonical form known as "disparate treatment," and it is our emphasis here. In some legal systems, discrimination is also said to occur when screeners adopt a practice that has a disproportionate negative effect on members of protected groups, at least when that practice does not have a powerful justification. Our discussion will bear on that understanding of discrimination as well; it is usually described as "disparate impact."

The Sources of Algorithmic Discrimination

Algorithms can be powerful tools that help people use data to accomplish their objectives more effectively. Sometimes the objective humans have in mind is to discriminate on some

[‡]There is an extensive literature on these dual-systems models; see ref. 9 for an influential overview, ref. 10 for implications in economics, and ref. 11 for implications for impulse control.

[§]See ref. 13 for connections between implicit bias and law and refs. 14 and 15 for the question of whether implicit attitudes map onto behavior.

[¶]Barocas and Selbst (4) provide an extensive discussion of the current and potential roles that algorithms have in these domains.

forbidden ground. This means that algorithms in principle have the potential to help humans discriminate more effectively and insidiously. It is critically important to ensure that we have the right laws and regulations in place to prevent impermissible discrimination.

Some terminology might first be useful. People often refer to any process that takes data and produces a subsequent prediction as an "algorithm." But it is important to note that there are actually two separate "algorithms" at work in screening applications of the type we are considering: 1) The screening algorithm (or screener) simply takes the characteristics of an individual (like a job applicant) and returns a prediction of this individual's outcome. This prediction then informs a decision (such as hiring). 2) The training algorithm (or trainer) is what produces the screening algorithm. Constructing the training algorithm involves (among other things) assembling past instances to use as training data, defining the outcome to predict, and choosing candidate predictors to consider.

The screening algorithm is just the mechanical result of running the training algorithm on a set of training data. So while the screening algorithm can produce biased decisions, the place where this algorithmic discrimination gets introduced must come from the human decisions that go into building the training algorithm.

In ref. 6, we showed formally that algorithmic bias can be decomposed completely into three components: bias in the choice of input variables, bias in the choice of outcome measure, and bias in the construction of the training procedure. After accounting for these three forms of bias, any remaining disparity corresponds to the structural disadvantage of one group relative to another.

To see this, consider a screening problem (e.g., hiring) with applicants from two distinct groups: an advantaged group and a disadvantaged group. Suppose that each applicant is described by a feature vector x, and the applicant's true productivity for the task at hand (the focus of the screening decision) is a function $f(x)$ of this vector. Let the fraction of applications with feature vector x be $p(x)$ in the advantaged group and $q(x)$ in the disadvantaged group. Average productivity of applicants in the advantaged and disadvantaged groups will equal $\sum_x p(x)f(x)$ and $\sum_x q(x)f(x)$, respectively. Finally, for an arbitrary function v, let $D(v)$ be the difference in the average value of v between the advantaged and disadvantaged groups; that is, $D(v) = \sum_x [p(x) - q(x)]v(x)$. Applying this notation to the function f, we see that $D(f)$ is the difference in average productivity between the two groups; this constitutes the structural disadvantage of one group relative to the other.

The designers of the algorithm know neither the true function f nor the applicant's full feature vector x.[#] So the process of building an algorithm proceeds as follows: The designer specifies applicant performance as $g(x)$, which is usually not exactly $f(x)$. Since the full feature x is generally not known, the designer constructs the algorithm based on a reduced feature vector $r(x)$. And since the function g is designed to be applied to vectors whose dimension is the same as x, a different function h must be used on the reduced vector $r(x)$. The result is a value $h(r(x))$. If we use $h \circ r$ to denote the composition of functions h and r (obtained by first applying r and then applying h), then the value for an applicant with feature vector x is $(h \circ r)(x)$. Finally, the function h must be estimated from the available training data; this estimate is a function t. The function t is applied to the reduced feature vector $r(x)$, resulting in the value $(t \circ r)(x)$. Thus, the design process results in an algorithm that takes an applicant with feature vector x and produces a score $(t \circ r)(x)$. The screening process ranks applicants by this score and selects the highest-ranked ones.

If we wish to understand the disparities between the advantaged and disadvantaged groups under the scores produced by the algorithm, we should look at the quantity $D(t \circ r)$. A central question in evaluating the possible biases introduced by the algorithm design process is to compare this disparity $D(t \circ r)$ with the underlying structural disadvantage $D(f)$.

A useful way to perform this comparison is to write $D(t \circ r)$ in the following extended form:

$$D(t \circ r) = D(f) + (D(g) - D(r)) + (D(h \circ r) - D(g)) + (D(t \circ r) - D(h \circ r)).$$

The terms on the right-hand side can then be interpreted as follows: The first term, $D(f)$, is the underlying structural disadvantage that was present prior to the construction of the algorithm. The second term, $D(g) - D(f)$, is the bias added by using g as an outcome measure instead of f. The third term, $D(h \circ r) - D(g)$, is the bias added by using $r(x)$ as the feature vector instead of x. The fourth term, $D(t \circ r) - D(h \circ r)$, is the bias added by the fact that we are using an estimated function t rather than the function h.

This makes clear that the human decisions that go into building an algorithm can, either intentionally or inadvertently, produce discrimination. Moreover, the existing legal systems that most countries have to detect and hence prevent bias are not well suited to uncover the specific forms of bias that humans introduce into algorithms. Consider that to hire people, a company needs only its screening algorithm. Once the screening algorithm has been produced, the company, in principle, no longer needs the training algorithm or training data, so a discriminating firm could potentially choose not to store them for strategic reasons (which would not be prohibited currently in the legal systems of most countries). In this sense, algorithms create new risks for discrimination absent any changes to existing laws and regulations.

On the other hand, the second, third, and fourth terms in this equation can be quantified in ways that bias in purely human-driven decision systems can never be. There is no way to know the true functions f, p, or q. But if the right laws and regulations are in place that allow regulators or relevant third parties to reanalyze the data and interrogate the choices made during the training stage, it is possible to determine the sensitivity of decision outcomes (including disparities) to the choice of outcome, candidate predictors, or machine-learning engineering procedures.[‖] Our point is not that discriminating firms or other organizations would voluntarily stop discriminating because of an algorithm, but rather that the use of an algorithm will now make detection easier by those with the authority to prevent them from discriminating.

In principle (and we emphasize those cautionary words), algorithms therefore have the potential to become a force for social justice by serving as powerful detectors of human discrimination. And as we have noted, algorithms can also help clarify, in an additional way, the normative issues raised by the very concept of discrimination, which can be understood in diverse ways by the legal system and which is often used loosely in public debates.

[#] That the algorithm builder does not know the true function f is one key reason why there are market returns to machine-learning engineering skill; variation in that skill is reflected, for example, in prediction competitions such as the "Netflix prize" or those organized by Kaggle. The possibility that the algorithm builder does not have access to the full set of information about people is discussed extensively in ref. 17.

[‖] For a more extensive discussion of what sorts of laws and regulations would help promote algorithmic fairness see ref. 6. For a discussion of how facially neutral algorithm construction procedures can nonetheless lead to unnecessarily large disparities in decision outcomes, see ref. 18.

Returning to Our Example

Suppose now that the tech firm in our hiring example from above had used an algorithm to screen applicants, one constructed via a machine-learning framework of training and screening algorithms. And let us also suppose, and this is crucial, that all this is happening at a time and place with a different kind of legal framework from what most countries have now. Suppose the law requires firms to store the screener and trainer code and training dataset and make them available in response to discrimination lawsuits. A reason for that framework should now be obvious. With it, it will be possible to know whether discrimination has occurred (and to get clearer on what discrimination is). Without it, that will not be possible.

The transparency and specificity of algorithms now create radically different opportunities to uncover discrimination. Rather than asking humans unanswerable questions, we have a clearer set of targets for inquiry that we can answer more precisely.

For starters, the fact that the plaintiff could access the screener algorithm means that it becomes possible to probe and experiment with it in ways that are not feasible with a human hiring process. It is true that algorithms are not designed to isolate the contribution of a particular individual applicant characteristic on the outcome, which is a difficult statistical task in high-dimensional applications where applicant characteristics (or any candidate predictor variables) are typically correlated with one another. But we can feed any set of applicant characteristics into the screener to answer counterfactual questions that humans could never answer, like "Would this specific candidate have been hired were the candidate a man instead of a woman?"

Suppose that in our hypothetical scenario, the plaintiff's statistical expert finds that changing applicant gender does not change hiring outcomes (suggesting that disparate treatment has not occurred), but that changing the client's other qualifications to those of the average male would (raising the question of whether discrimination, in some intelligible form, may nonetheless have taken place). The statistical expert can explore why: Is that result because of the firm's choice of training procedure, the choice of inputs, or the choice of outcome? Is any one of these choices objectionable?

One straightforward way the plaintiff's statistical consultants can detect at least some form of discrimination in the training procedure is to build their own training algorithm from scratch and compare the results to those of the firm's actual algorithm. This can help illuminate the question of whether there has been some kind of discrimination in a way that reading the trainer's code cannot. Suppose the statistical experts find they cannot build a model as predictive as the firm's with smaller disparities; this rules out concern that the firm underinvested in machine-learning engineering or even hid some bias in the code. The expert recognizes that another way the trainer could be contaminated is if the firm used a biased sample of training data. But the expert finds that the training dataset is statistically representative of the universe of past applicants and hired workers at the firm and of the larger population and that the correlation of features and labels is similar to what is seen in other datasets. This helps rule out things like the possibility that the firm "gamed" its training dataset, for example to create a bad track record for women by considering only less productive female applicants for a period (19).

The plaintiff's expert can also now check for discrimination with the candidate predictors or features by comparing what the firm actually used to the full set of variables the firm has available and to what other firms in the same industry use. The motivation for this industry benchmark comes from evidence that while discrimination remains a key problem, not all firms discriminate (20).

Finally the expert can now also carry out a series of tests for arguable forms of discrimination in the firm's choice of label or outcome to predict. The label being predicted is something that must be explicitly specified in the process of training, and hence in our scenario regulation can require it to be made available as part of litigation. Suppose it turns out (in our hypothetical example) that the outcome chosen by the firm leads to much larger gender disparities in hiring than many other plausible candidate outcomes the firm could have used. By itself, this may or may not provide sufficient evidence of discrimination in court; everything depends on the legal standard. But if the firm realizes that it lacks a good explanation for its choice of outcome to predict, it might change its practice.

We have, in sum, gone from a scenario with purely human-driven decision making where getting any useful insights is difficult to one where, if the right laws and regulations are in place, we have opportunities to carry out concrete tests for each of the ways in which humans might introduce discrimination. Perhaps it goes without saying that there is no guarantee that society will converge to the right laws and regulations for algorithms, given the difficulty of getting the right regulations in place for a wide range of human behaviors and given normative disputes about what should be counted as discrimination.** Nor do we claim the specific tests we have outlined above will ultimately prove to be the best ones or that foolproof tests will ever be found. Our argument instead is that the use of algorithms creates the potential to reduce discrimination (however it is understood) relative to the status quo and hence creates very high social returns to efforts that realize this potential by changing laws and regulations and developing tests for algorithmic bias.

Scaling Solutions

Detection is a necessary but not sufficient condition for preventing discrimination. Prevention is particularly difficult in settings that involve entirely human-driven decision loops, as evidenced for example by the persistence of discrimination in settings where it has been previously documented. Beyond making it much easier to detect discrimination, the introduction of an algorithm into the decision loop now makes it also much more feasible to identify and scale useful fixes.

To see this, imagine that in our hypothetical case study the litigation did somehow manage to turn up evidence of discrimination, understood as disparate treatment. Suppose, for instance, that an HR manager forgot to delete a text off a phone that overtly discussed plans to engage in discriminatory behavior, and this gets discovered as part of litigation. What then? The plaintiff might get a settlement or win a lawsuit, but this might not be enough to deter future human discrimination because this type of smoking-gun evidence is rare and firms will take steps to avoid it.

The plaintiff might seek, as part of a decree or settlement, an agreement from the firm to change its hiring practices as well. But in a purely human-driven hiring system, what should the people running the firm actually do in practice? Suppose they instructed their HR team not to discriminate. Given problems of explicit and implicit bias, and the general challenges of detecting human discrimination, how could the firm's leadership be sure this was doing any good?

Contrast this with the case where an algorithm is now involved in hiring. Returning to our example from above, suppose the plaintiff's statistical consultant does a bit more digging into the

**For example, taxes in the United States on alcohol remain far too low given the external costs that drinking imposes on society (21); regulations on human activities that contribute to climate change remain far from ideal (see, for example, ref. 22); and the global financial crisis of 2008 highlighted a number of important limitations of existing market regulations (for example, ref. 23).

source of the disparities in hiring outcomes. The answer is the algorithm is built to predict the past hiring decisions of the firm's human HR managers, which are themselves discriminatory. Now there is a clear and scalable solution: Swap in a different outcome to be predicted that is less infected by human discrimination, as revealed for instance by smaller disparities in hiring decisions. One obvious candidate, for instance, is actual job performance (such as coding skill, at a tech firm), given studies suggesting that biases in subjective evaluation do not correspond to differences in this measure (24).[††]

Conclusion

It is tempting to think that human decision making is transparent and that algorithms are opaque. But as we have argued here, with respect to discrimination the opposite is true—or could be true, if we put the right laws and regulations in place to capitalize on the far greater specificity and transparency that algorithms make possible. That is an essential task. We note that here we have emphasized its importance while only hinting about how, exactly, to carry it out.

We have argued that detecting human bias in purely human-driven decision processes is enormously difficult. The most convincing empirical support we can marshal for this claim comes from the sheer prevalence of human bias in practice. We know this from audit studies, which cannot identify when a specific person has discriminated, but can tell us something about how common bias is overall. For example, in the United States, audit studies that randomly assign otherwise-equivalent White and Black applicants to apply at different firms find that White applicants are called back at more than twice the rate of Black applicants, 34% versus 14% (1). Reducing this discrimination would do an enormous amount of social good given there are over 6 million job openings in the United States at any point in time (25). Audit studies of the US housing market, which originates $2 trillion in new mortgages each year (26), find that minority borrowers are treated differently from and worse than White borrowers (27). Discrimination also arises in the health sector, which accounts for $3.5 trillion in spending each year in the United States alone (equal to 18% of gross domestic product) (28, 29). For example, when doctors were shown two equivalent patient histories, the chances of recommending a beneficial procedure (cardiac catheterization) were 40% lower for women and minorities than for White males (30).

Of course, in the real world algorithms will for the foreseeable future understandably (and perhaps appropriately) be used as decision aids to help humans, rather than decision makers that supplant humans. Nothing about our argument fundamentally changes in this case. We can still learn about the algorithmic part of the decision-making system in ways that we never could with purely human-driven systems. Introducing a data-driven decision aid creates additional opportunities to detect what the humans in the system are doing, since we can test whether human compliance with the tool's recommendations, as opposed to override, is systematically lower or higher for protected groups. And to the extent that the algorithm's recommendations create something like a default for the decision, they may create opportunities to "nudge" human decisions in the direction of reducing discrimination however it is understood, including disparities in decision outcomes (31).

The risk that algorithms introduce is not from their use per se, but rather the risk that our regulatory and legal systems will not keep pace with the changing technology. But if we make the necessary adjustments to account for the different world we are in, algorithms have enormous potential to be not just a risk to be managed but actually a force for social good. The use of algorithms offers far greater clarity and transparency about the ingredients and motivations of decisions. At the conceptual level, this provides unprecedented opportunities to understand what, exactly, discrimination is and thus to achieve increased clarity on some of the most contested normative issues in contemporary societies. And more pragmatically, it provides powerful opportunities to detect, and hence to help prevent, discrimination in many places where it may occur.

Data Availability. This work discusses policy questions associated with algorithms for screening decisions; there are no data specifically associated with this paper.

[††]In some jobs not every outcome that the decision maker cares about will be quantifiable. But given evidence from behavioral science that human biases are particularly pronounced in unstructured decision-making environments, the more the criteria used to inform screening decisions can be quantifiable (and hence predicted with data-driven decision-making tools), the greater the potential to reduce discrimination in practice.

1. D. Pager, The mark of a criminal record. *Am. J. Sociol.* **108**, 937–975 (2003).
2. M. Bertrand, S. Mullainathan, Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
3. L. Sweeney, Discrimination in online ad delivery. *Queue* **11**, 10–29 (2013).
4. S. Barocas, A. Selbst, Big data's disparate impact. *Calif. Law Rev.* **104**, 671–732 (2016).
5. S. Bornstein, Antidiscriminatory algorithms. *Ala. Law Rev.* **70**, 519–572 (2018).
6. J. Kleinberg, J. Ludwig, S. Mullainathan, C. Sunstein, Discrimination in the age of algorithms. *J. Legal Anal.* **10**, 113–174 (2018).
7. G. Stoet, D. Geary., The gender-equity paradox in science, technology, engineering and mathematics education. *Psychol. Sci.* **29**, 581–593 (2018).
8. T. Wilson, *Strangers to Ourselves* (Harvard University Press, Cambridge, MA, 2004).
9. D. Kahneman, *Thinking Fast and Slow* (Farrar, Straus and Giroux, New York, NY, 2011).
10. T. Cunningham, Biases and implicit knowledge. https://www.parisschoolofeconomics.eu/IMG/pdf/cunningham_paper.pdf. Accessed 9 July 2020.
11. D. Fudenberg, D. Levine, A dual-self model of impulse control. *Am. Econ. Rev.* **96**, 1449–1476 (2006).
12. D. Kahneman, *Attention and Effort* (Prentice-Hall, Englewood Cliffs, NJ, 1973).
13. C. Jolls, C. Sunstein, The law of implicit bias. *Calif. Law Rev.* **94**, 969–996 (2006).
14. H. Arkes, P. Tetlock, Attributions of implicit prejudice. *Psychol. Inq.* **15**, 257–278 (2004).
15. P. Tetlock, G. Mitchell, Implicit bias and accountability systems. *Res. Organ. Behav.* **29**, 3–38 (2009).
16. L. Cosmides *et al.*, Perceptions of race. *Trends Cognit. Sci.* **7**, 173–179 (2003).
17. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).
18. M. Hu, Algorithmic Jim Crow. *Fordham Law Rev.* **86**, 633–696 (2017).
19. C. Dwork *et al.*, "Fairness through awareness" in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, S. Goldwasser, Ed. (ACM Press, New York, NY, pp. 214–226.
20. K. Charles, J. Guryan, Prejudice and wages: An empirical assessment of Becker's the Economics of Discrimination. *J. Polit. Econ.* **116**, 773–809 (2008).
21. P. Cook, *Paying the Tab: The Costs and Benefits of Alcohol Control* (Princeton University Press, Princeton, NJ, 2007).
22. N. Stern, J. Stiglitz, *Report of the High-Level Commission on Carbon Prices* (Carbon Pricing Leadership Coalition, 2017).
23. A. Taylor, Credit, financial stability, and the macroeconomy. *Annu. Rev. Econ.* **7**, 309–339 (2015).
24. J. Terrell *et al.*, Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Comput. Sci.* **3**, e111 (2017).
25. Bureau of Labor Statistics, Job openings and labor turnover summary (Economic News Release). https://www.bls.gov/news.release/jolts.nr0.htm. Accessed 8 January 2019.
26. T. Kapfidze, U.S. mortgage market statistics: 2018. Magnify money by lending tree. https://www.magnifymoney.com/blog/mortgage/u-s-mortgage-market-statistics-2017/. Accessed 21 December 2018.
27. M. Turner *et al.*, "All other things being equal: A paired testing study of mortgage lending institutions–final report" (Tech. Rep., US Department of Housing and Urban Development OFfice of Policy Development and Research, Washington, DC, 2002).
28. Center for Medicare & Medicaid Services, Health expenditure data for 2017. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html. Accessed 9 July 2020.
29. World Bank, GDP. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD. Accessed 9 July 2020.
30. K. Schulman *et al.*, The effect of race and sex on physicians' recommendations for cardiac catheterization. *N. Engl. J. Med.* **340**, 618–626 (1999).
31. R. Thaler, C. Sunstein, *Nudge: Improving Decisions about Health, Wealth and Happiness* (Yale University Press, New Haven, CT, 2008).