Algorithms, Correcting Biases

Cass R. Sunstein

social research

AN INTERNATIONAL QUARTERLY

ALGORITHMS

Social Research: An International Quarterly, Volume 86, Number 2, Summer 2019, pp. 499-511 (Article)

Published by Johns Hopkins University Press

Cass R. Sunstein

Algorithms, Correcting Biases

ALGORITHMS, BAIL, AND JAIL

Are algorithms biased? If so, in what respect? These are large questions, and there are no simple answers. My goal here is to offer one perspective on them, principally by reference to some of the most important current research on the use of algorithms for purposes of public policy and law. I offer two claims. The first, and the simpler, is that algorithms can overcome the harmful effects of cognitive biases, which can have a strong hold on people whose job it is to avoid them and whose training and experience might be expected to allow them to do so. Many social questions present prediction problems, where cognitive biases can lead people astray; algorithms can serve as a corrective (see Kleinberg et al. 2015, 105).

In a way, this should be an unsurprising claim. Some of the oldest and most influential work in behavioral science shows that statistical prediction often outperforms clinical prediction; one reason involves cognitive biases on the part of clinicians (Meehl [1953] 2013). Algorithms can be seen as a modern form of statistical prediction, and if they avoid biases, no one should be amazed. What I hope to add here is a concrete demonstration of this point in an important context, with some general remarks designed to address the concern that algorithms are biased.

The second claim, and the more complex one, is that algorithms can be designed so as to avoid racial (or other) discrimination in its unlawful forms—and also raise hard questions about how to balance competing social values (Kleinberg et al. 2019). When people complain about algorithmic bias, they are often concerned about discrimination on the basis of race and sex. The word "discrimination" can be understood in many different ways. It should be simple to ensure that algorithms do not discriminate in the way that American law most squarely addresses. It is less simple to deal with forms of inequality that concern many people, including an absence of "racial balance." As we shall see, algorithms allow new transparency about some difficult tradeoffs (Kleinberg et al. 2018; Kleinberg et al. 2019).

The principal research on which I focus comes from Jon Kleinberg, Himabinku Lakkaraju, Jure Leskovec, Jens Lutwig, and Sendhil Mullainathan, who explore judges' decisions about whether to release criminal defendants pending trial (Kleinberg et al. 2018, 237). Their goal is to compare the performance of an algorithm with that of actual human judges, with particular emphasis on the solution to prediction problems. It should be obvious that the decision about whether to release defendants has large consequences. If defendants are incarcerated, the long-term consequences could be very severe. Their lives could be ruined, or nearly so. But if defendants are released, they might flee the jurisdiction or commit crimes. People might be assaulted, raped, or killed.

In some states, the decision whether to allow pretrial release turns on a single factor: flight risk. To make their decision, judges have to solve a prediction problem: *What is the likelihood that a defendant will flee the jurisdiction*? In other states, the likelihood of crime also matters, and it too presents a prediction problem: *What is the likelihood that a defendant will commit a crime*? (As it turns out, flight risk and crime are closely correlated, so that if one accurately predicts the first, one would accurately predict the second as well.) Kleinberg and his colleagues built an algorithm that uses, as inputs, the same data available to judges at the time of a bail hearing, such as prior criminal history and current offense. Their central finding is that *along every dimension that matters, the algorithm does much better than real-world judges*. Among other things:

1. Use of the algorithm could maintain the same detention rate now produced by human judges and reduce crime by up to 24.7 percent. Alternatively, use of the algorithm could maintain the current level of crime reduction and reduce jail rates by as much as 41.9 percent. That means that if the algorithm were used instead of judges, thousands of crimes could be prevented without jailing even one additional person. Alternatively, thousands of people could be released, pending trial, without adding to the crime rate. It should be clear that use of the algorithm would allow any number of political choices about how to balance decreases in the crime rate against decreases in the detention rate.

2. A major mistake made by human judges is that they release many people identified by the algorithm as especially high-risk (meaning likely to flee or to commit crimes). More specifically, judges release 48.5 percent of the defendants judged by the algorithm to fall in the riskiest 1 percent. Those defendants fail to reappear in court 56.3 percent of the time. They are rearrested at a rate of 62.7 percent. Judges show leniency to a population that is likely to commit crimes.

3. Some judges are especially strict, in the sense that they are especially reluctant to allow bail—but their strictness is not limited to riskiest defendants. If it were, the strictest judges could jail as many people as they now do, but with a 75.8 percent increase in reduction of crime. Alternatively, they could keep the current crime reduction, and jail only 48.2 percent as many people as they now do.

A full account of why the algorithm outperforms judges would require an elaborate treatment. But for my purposes here, a central part of the explanation is particularly revealing. As point 2 above suggests, judges do poorly with the highest-risk cases. (This point holds for the whole population of judges, not merely for those who are the strictest.) The reason is an identifiable bias (Kleinberg et al. 2018, 284); call it *current offense bias*.

On this count, Kleinberg and his colleagues restrict their analysis to two brief sentences, but those sentences have immense importance (284). As it turns out, judges make two fundamental mistakes. First, they treat high-risk defendants as if they are low-risk *when their current charge is relatively minor* (for example, it may be a misdemeanor). Second, they treat low-risk people as if they are high-risk *when their current charge is especially serious*. The algorithm makes neither mistake. It gives the current charge its appropriate weight. It takes that charge in the context of other relevant features of the defendant's background, neither overweighting nor underweighting it. The fact that judges release a number of high-risk defendants is attributable, in large part, to overweighting the current charge (when it is not especially serious).

AVAILABILITY BIAS AND ITS COUSINS

Current offense bias is of particular interest. It shows that when human beings suffer from a cognitive bias, a well-designed algorithm, attempting to solve a prediction problem, can do much better. It is worth emphasizing that we are dealing not with novices, but with human beings who are both trained and experienced. They are experts. Nonetheless, they suffer from a cognitive bias that produces severe and systematic errors.

The point has more general interest. Current offense bias is best understood as a close cousin of *availability bias*: individual judgments about probability are frequently based on whether relevant examples are easily brought to mind (Tversky and Kahneman 1982, 3). Both biases involve *attribute substitution* (Kahneman and Frederick 2002; Kahneman 2011). Availability bias is a product of the availability heuristic, which people use to solve prediction problems. They substitute a relatively easy question ("does an example come to mind?") for a difficult one ("what is the statistical fact?"). Current offense bias reflects what we might call the *current offense heuristic*, which also involves an easy question.

Because of the availability heuristic, people are likely to think that more words, on a random page, end with the letters "ing" than have "n" as their next to last letter (Tversky and Kahneman 1982, 3)—even though a moment's reflection will show that this could not possibly be the case. Furthermore, "a class whose instances are easily retrieved will appear more numerous than a class of equal frequency whose instances are less retrievable" (11). Consider a simple study showing participants a list of well-known people of both sexes and asking them whether the list contains more names of women or more names of men. For lists in which the men were especially famous, people thought that there were more names of men, whereas for lists in which the women were the more famous, people thought that there were more names of women (11).

This is a point about how *familiarity* can affect the availability of instances and thus produce mistaken solutions to prediction problems. A risk that is familiar, like that associated with smoking, will be seen as more serious than a risk that is less familiar, like that associated with sunbathing. But *salience* is important as well. "For example, the impact of seeing a house burning on the subjective probability of such accidents is probably greater than the impact of reading about a fire in the local paper" (11). *Recency* matters as well. Because recent events tend to be more easily recalled, they have a disproportionate effect on probability judgments. Availability bias thus helps account for "recency bias" (Ashton and Kennedy 2002, 221). Current offense bias can be understood as a sibling to recency bias.

In many domains, public officials and private citizens must solve prediction problems, and availability bias can lead to damaging and costly mistakes. Whether people will buy insurance for natural disasters is greatly affected by recent experiences (Slovic 2000, 40). If floods have not occurred in the immediate past, people who live on flood plains are far less likely to purchase insurance. In the aftermath of an earthquake, insurance for earthquakes rises sharply—but it declines steadily from that point, as vivid memories recede. Note that the use of the availability heuristic, in these contexts, is hardly irrational. Both insurance and precautionary measures can be expensive, and what has happened before seems, much of the time, to be the best available guide to what will happen in the future. The problem is that the availability heuristic can lead to serious errors, stemming from both excessive fear and neglect.

If the goal is to make accurate predictions, use of algorithms can be a great boon for that reason. For both private and public institutions (including governments around the world), it can eliminate the effects of cognitive biases. Suppose that the question is whether to hire a job applicant; or whether a project will be completed within six months; or whether a particular intervention will help a patient who suffers from heart disease. In all these cases, some kind of cognitive bias may well distort human decisions. There is a good chance that availability bias or one of its cousins will play a large role; and unrealistic optimism, embodied in the planning fallacy, may aggravate the problem. Algorithms have extraordinary promise. They can save both money and lives.

DISCRIMINATION

In the current period, there is a great deal of concern that algorithms discriminate on illegitimate grounds, such as race or sex (Chander 2017, 1023). The concern appears to be growing. The possibility that algorithms would increase or promote discrimination raises an assortment of difficult questions. But the bail research casts new light on them. Above all, it suggests a powerful and simple point: use of algorithms will reveal, with great clarity, the need to make tradeoffs between the value of racial (or other) equality and other important values, such as public safety.

A (Very) Little Law

Discrimination law has long been focused on two different problems. The first is *disparate treatment*; the second is *disparate impact* (Barocas and Selbst 2016, 672). The Constitution and all civil rights laws forbid disparate treatment. The Constitution does not concern itself with disparate impact, but some civil rights statutes do.

1. *Disparate treatment*. The prohibition on disparate treatment reflects a commitment to a kind of neutrality. Public officials are not permitted to favor members of specified groups over others unless there is a sufficiently neutral reason for doing so. Different civil rights laws forbid disparate treatment along a variety of specified grounds, such as race, sex, religion, disability, and age. In extreme cases, the existence of disparate treatment is obvious, because a facially discriminatory practice or rule can be shown to be in place ("no women may apply"). In other cases, no such practice or rule can be identified, and for that reason, violations are more difficult to police. A plaintiff might claim that a facially neutral practice or requirement (such as a written test for employment) was actually adopted in order to favor one group (whites) or to disfavor another (African Americans). To police discrimination, the legal system is required to use what tools it has to discern the motivation of decision makers.

Such violations might arise because of explicit prejudice, sometimes described as "animus." Alternatively, they might arise because of unconscious prejudice, operating outside the awareness of the decision maker; unconscious prejudice is sometimes described as an "implicit bias." An official might discriminate against women not because he intends to do so but because of an automatic preference for men, which he might not acknowledge and might even deplore.

2. *Disparate impact*. The prohibition on disparate impact means, in brief, that if some requirement or practice has a disproportionate adverse effect on members of specified groups (African Americans, women), the manager must show that it is adequately justified. Suppose, for example, that an employer requires members of its sales force to take some kind of written examination, or that the head of

a police department institutes a rule requiring new employees to be able to run at a specified speed. If these practices have disproportionate adverse effects on African Americans or women, they would be invalidated unless they could show a strong connection to the actual requirements of the job. In many cases, they must show that the practices are justified by "business necessity."

The theory behind disparate impact remains disputed. On one view, the goal is to ferret out disparate treatment. If, for example, an employer has adopted a practice with disproportionate adverse effects on African Americans, we might suspect that it is intending to produce those adverse effects. The required justification is a way of seeing whether the suspicion is justified. Alternatively, disparate impact might be thought to be disturbing in itself, in the sense that a practice that produces such an impact helps entrench something like a caste system. If so, it would be necessary for those who adopt such practices to demonstrate that they have a good and sufficiently neutral reason for doing so.

Algorithms, Judges, and Bail

Human judges

In the context of bail decisions, we would have disparate treatment if it could be shown that judges discriminated against African American defendants, either through a formal practice (counting race as a "minus") or through a demonstrable discriminatory motive (established perhaps with some kind of extrinsic evidence). We would have disparate impact if it could be shown that some factor or rule of decision (taking account, for example, of employment history) had a disproportionate adverse effect on African Americans; the question would be whether that effect could be adequately justified in neutral terms.

For present purposes, let us simply assume that the decisions of human judges, with respect to bail decisions, show neither disparate treatment nor disparate impact. As far as I am aware, there is no proof of either in the relevant data (that is, there is no proof that human judges are discriminating in either respect, as a matter of law). It is nonetheless true that for African Americans and Hispanics, the detention rate is 28.5 percent (Barocas and Selbst 2016, 672). More specifically, African Americans are detained at a rate of 31 percent, and Hispanics are detained at a rate of 25 percent. (The detention rate for whites is between those two figures.) Because the class of criminal defendants is disproportionately African American and Hispanic (whites constitute less than 13 percent), the group of people who are denied bail is disproportionately African American and Hispanic as well.

The algorithm

Importantly, the algorithm is made blind to race. Whether a defendant is African American or Hispanic is not one of the factors that it considers in assessing flight risk. But with respect to *outcomes*, how does the algorithm compare to human judges?

The answer, of course, depends on what the algorithm is asked to do. If the algorithm is asked to achieve the same *crime rate* as the judges, it could do that by jailing 38.8 percent fewer African Americans and 44.6 percent fewer Hispanics. (It can do that because the pool of defendants is disproportionately nonwhite, and so any effort to cut the detention rate, while maintaining the same crime rate, will have disproportionate benefits for nonwhites.) If the algorithm is directed to match the judges' overall *detention rate*, its numbers, with respect to race, would look quite close to the corresponding numbers for those judges. Its overall detention rate for African Americans or Hispanics is 29 percent, with a 32 percent rate for African Americans and 24 percent for Hispanics. At the same time, the crime rate drops, relative to judges, by a whopping 25 percent.

It would be fair to say that on any view, the algorithm is not a discriminator, at least not when compared with human judges. There is no disparate treatment. It would be difficult to find disparate impact. And in terms of outcomes, it is not worse along the dimension of racial fairness. (Whether the numbers are nonetheless objectionable is a separate question.)

Kleinberg et al. show that it is also possible to constrain the algorithm to see what happens if we aim to reduce that 29 percent detention rate for African Americans and Hispanics (Kleinberg et al. 2018). Suppose that the algorithm is constrained so that the detention rate for African Americans and Hispanics has to stay at 28.5 percent. It turns out that the crime reduction is about the same as would be obtained with the 29 percent rate. Moreover, it would be possible to instruct the algorithm in multiple different ways, so as to produce different tradeoffs among social goals. The authors give some illustrations: *maintain the same detention rate but equalize the release rate for all races*. The result is that the algorithm reduces the crime rate by 23 percent—significantly but not massively lower than the 25 percent rate achieved without the instruction to equalize the release rate. And recall that if the algorithm is instructed to produce the same crime rate that judges currently achieve, it would jail many fewer African Americans and Hispanics, because it would detain many fewer people, focusing on the riskiest defendants; many African Americans and Hispanics would benefit from its more accurate judgments.

The most important point here may not involve the particular numbers, but instead the clarity of the tradeoffs. The algorithm would permit any number of choices with respect to the racial composition of the population of defendants denied bail. It would also make explicit the consequences of those choices for the crime rate.

BROADER CONSIDERATIONS

When it is said that algorithms can correct for biases, what is usually meant is cognitive biases (such as current offense bias). The case of discrimination is more challenging. To be sure, disparate treatment can usually be prevented; algorithms do not have motivations, and they can be designed so as not to draw lines on the basis of race or sex, or to take race or sex into account. The case of disparate impact is trickier. If the goal is accurate predictions, an algorithm might use a factor that is genuinely predictive of what matters (flight risk, educational attainment, job performance)—but that factor might have a disparate impact on African Americans or women. If disparate impact is best understood as an effort to ferret out disparate treatment, it might not be a problem, at least so long as no human being, armed with a discriminatory motive, is behind its use, or behind the factor that is being used. But the "so long as" qualification is important. And if disparate impact is an effort to prevent something like a caste system, it might deserve scrutiny.

Especially difficult problems are presented if an algorithm uses a factor that is in some sense an outgrowth of discrimination. For example, a poor credit rating or a troubling arrest record might be an artifact of discrimination by human beings that occurred before the algorithm was asked to do its predictive work. There is a risk here that algorithms could perpetuate discrimination and extend its reach, by using factors that are genuinely predictive but products of unequal treatment (Kleinberg et al. 2019). This might turn discrimination into a kind of self-fulfilling prophecy.

In terms of existing law, racial balance, as such, is not legally mandated, and efforts to pursue that goal might themselves be struck down on constitutional grounds. Nonetheless, some people are keenly interested in reducing racial and other disparities—for example, in education, in health care, and in the criminal justice system. One of the signal virtues of algorithms is that they present the relevant trade-offs in an unprecedentedly clear light. We might learn that if we pursue racial balance, we would sacrifice other goals, and we might be able to see, with real precision, the magnitude of the gains and the losses. One advantage of the bail study is that it offers a clear illustration. The tradeoffs might well be painful, but in general, it is best to know what they are.

BEYOND INTUITIONS

The use of algorithms is often motivated by an appreciation of the limitations of human intuition. In the private and public sectors, people are often asked to make predictions under conditions of uncertainty, and their intuitions can lead them astray (Hertwig 2019). It takes a great deal of work to provide corrections (Beyth-Marom and Dekel [1980] 2010). It is often believed that experts can develop reliable intuitions, or rely instead on statistical thinking. That is frequently true, at least when they receive prompt feedback. But as current offense bias makes clear, experienced judges (in the literal sense) can do significantly worse than algorithms. Antagonism toward algorithms is often based, I suggest, on fallible intuitions, though the full story would require extended elaboration and many qualifications.

There is no assurance, of course, that algorithms will avoid cognitive biases. They can be built so as to display them. My point is that they can also be built so as to improve on human decisions. This is simply a specification of the old finding that statistical prediction often outperforms clinical prediction.

The problem of discrimination is different and far more complex, and I have only scratched the surface here, with reference to one set of findings. It is important to distinguish between (1) disparate treatment and (2) disparate impact, and it is also important to give separate treatment to (3) efforts to ensure that past discrimination is not used as a basis for further discrimination and (4) efforts to ensure racial or gender balance. For the future, (2) and (3) will present many of the most important issues for the use of algorithms. For (4), a primary advantage of algorithms is unprecedented transparency: they force people to make judgments about tradeoffs among compelling but sometimes incompatible policy goals.

ACKNOWLEDGEMENTS

I am grateful to Sendhil Mullainathan for many discussions; he deserves credit for anything that is useful here, and no blame for any of my mistakes.

REFERENCES

Ashton, Robert and Jane Kennedy. 2002. “Eliminating Recency with Self-Review: The Case of Auditors’ ‘Going Concern’ Judgments.” *Journal of Behavioral Decision Making* 15 (3): 221–31.

Barocas, Solon and Andrew D. Selbst. 2016. “Big Data’s Disparate Impact.” *California Law Review* 104 (671): 671–732.

Beyth-Marom, Ruth, Shlomith Dekel, Ruth Gombo, and Moshe Shaked. [1985] 2013. *An Elementary Approach to Thinking under Uncertainty.* Translated by Sarah Lichtenstein, Benny Marom, and Ruth Beyth-Marom. New York: Routledge.

Chander, Anupam. 2017. “The Racist Algorithm?” *Michigan Law Review* 115 (6): 1023–45.

Hertwig, Ralph, Timothy J. Pleskac, and Thorsten Pachur. 2019 forthcoming. *Taming Uncertainty.* Cambridge, MA: MIT Press.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus, and Giroux.

Kahneman, Daniel and Shane Frederick. 2002. “Representativeness Revisited: Attribute Substitution in Intuitive Judgment.” In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman. New York: Cambridge U. Press, 49–81.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. “Prediction Policy Problems.” *American Economic Review* 105 (5): 491–95.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions,” *Quarterly Journal of Economics* 133 (1): 237–93.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2019. “Discrimination in an Age of Algorithms.” *Journal of Legal Analysis* 10: 1–62.

Meehl, Paul. [1953] 2013. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Brattleboro, VT: Echo Point Books & Media.

Slovic, Paul. 2000. *The Perception of Risk: Risk, Society, and Policy.* New York: Earthscan.

Tversky, Amos and Daniel Kahneman. 1982. “Judgments of and by Representativeness,” in *Judgment under Uncertainty: Heuristics and Biases*. Edited by Daniel Kahneman, Paul Slovic, and Amos Tversky. New York: Cambridge U. Press, 84–100.