Duke Law Journal

VOLUME 71 MARCH 2022 NUMBER 6

# GOVERNING BY ALGORITHM? NO NOISE AND (POTENTIALLY) LESS BIAS

CASS R. SUNSTEIN†

## ABSTRACT

*As intuitive statisticians, human beings suffer from identifiable biases—cognitive and otherwise. Human beings can also be "noisy" in the sense that their judgments show unwanted variability. As a result, public institutions, including those that consist of administrative prosecutors and adjudicators, can be biased, noisy, or both. Both bias and noise produce errors. Algorithms eliminate noise, and that is important; to the extent that they do so, they prevent unequal treatment and reduce errors. In addition, algorithms do not use mental shortcuts; they rely on statistical predictors, which means that they can counteract or even eliminate cognitive biases. At the same time, the use of algorithms by administrative agencies raises many legitimate questions and doubts. Among other things, algorithms can encode or perpetuate discrimination, perhaps because their inputs are based on discrimination, or perhaps because what they are asked to predict is infected by discrimination. But if the goal is to eliminate discrimination, properly constructed algorithms nonetheless have a great deal of promise for administrative agencies.*

† Robert Walmsley University Professor, Harvard University. Parts of this Article draw on Cass R. Sunstein, *Algorithms, Correcting Biases*, 86 SOC. RSCH. 499 (2019), and I am grateful for permission to do that here. Special thanks as well to my coauthors, Daniel Kahneman and Olivier Sibony, with whom I have had many conversations on noise, bias, and algorithms; they are not responsible for my errors and confusions here. Special thanks too to Sendhil Mullainahan for many conversations on algorithms, their promise, and their limitations; the same proviso applies. Special thanks as well to Jens Ludwig for valuable comments on a previous draft. Eric Xu provided superb research assistance.

## TABLE OF CONTENTS

## INTRODUCTION

Should administrative agencies use algorithms? More than they do now? To what extent should the Internal Revenue Service use algorithms? The Environmental Protection Agency? The Transportation Security Administration? Ought we to move in the direction of an algorithmic state, or government by algorithm?[1] To answer these questions, we have to know something about the nature and magnitude of both bias and noise under processes that do and do not use algorithms. If algorithms reduce bias and noise, then there is strong reason to enlist them, even if that reason may not be inconclusive.[2] Agencies might want to enlist algorithms as advisers, on the ground that they provide relevant information. Alternatively, they

[1] *See* DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 6–8 (2020).

[2] On why it might not be conclusive, see DANIEL KAHNEMAN, OLIVIER SIBONY & CASS R. SUNSTEIN, NOISE: A FLAW IN HUMAN JUDGMENT 339–49 (2021), exploring reasons for individualized treatment, even if it results in noise.

might want to enlist algorithms as deciders, on the ground that they will do better than people will.

My principal claims here are threefold. *First*: Algorithms eliminate noise, and that is important; to the extent that they do so, they prevent unequal treatment and reduce errors. *Second*: Algorithms do not use mental shortcuts; they rely on statistical predictors, which means that they can counteract or even eliminate cognitive biases. *Third*: Algorithms can encode or perpetuate discrimination, perhaps because their inputs are based on discrimination or because what they (accurately) predict is infected by discrimination; but if the goal is to eliminate discrimination, properly constructed algorithms nonetheless hold a great deal of promise for the administrative state.

One of my starting points should be familiar. As intuitive statisticians solving prediction problems, human beings suffer from multiple biases.[3] We might show "availability bias," basing our judgments about probability on whether relevant examples are easily brought to mind.[4] We might be affected by "anchors," creating arbitrary numerical projections in light of them.[5] We might use the "affect heuristic," making judgments about products, proposals, people, and activities on the basis of our affective reactions to them, even if and when our judgments should be based on some kind of deliberation or even statistical analysis.[6] We might be unrealistically optimistic and thus show "optimistic bias," thinking that things will go better than they actually will.[7] As a result, we might fall prey to the planning fallacy, understood as the tendency to think that projects will take less long than they actually take.[8] We might display "present bias,"

[3] For an overview, see generally R.F. POHL, COGNITIVE ILLUSIONS (Pohl ed. 2016).

[4] Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, *in* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 3, 11 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982) [hereinafter Tversky & Kahneman, *Judgment Under Uncertainty*].

[5] An "anchor" is often understood as some numerical value, possibly provided at random, that affects numerical estimates. *See* Karen E. Jacowitz & Daniel Kahneman, *Measures of Anchoring in Estimation Tasks*, 21 PERSONALITY & SOC. PSYCH. BULL. 1161, 1161 (1995) (discussing how people who are presented with arbitrary values are more likely to make an estimate close to that number).

[6] *See* Paul Slovic, Melissa L. Finucane, Ellen Peters & Donald G. MacGregor, *The Affect Heuristic*, 177 EUR. J. OPERATIONAL RSCH. 1333, 1334 (2007).

[7] *See* TALI SHAROT, THE OPTIMISM BIAS 40 (2011).

[8] *See, e.g.*, DANIEL KAHNEMAN, THINKING, FAST AND SLOW 245–47 (2011) (anecdotally describing the planning fallacy); *see also* Roger Buehler, Dale Griffin & Michael Ross, *Exploring the "Planning Fallacy": Why People Underestimate Their Task Completion Times*, 67 J. PERSONALITY & SOC. PSYCH. 366, 366 (1994) (defining the planning fallacy). *See generally*

focusing on the short term and neglecting the long term.[9] Independently, we might be biased against various social groups, including people of color, women, people with disabilities, and the elderly, even if we are unaware of our biases.[10]

Human beings are also "noisy," regardless of whether we are biased.[11] Noise consists of unwanted variability in judgments; by contrast, bias consists of any systematic error that inclines people's judgments in a particular direction.[12] A bathroom scale might be noisy: it might show people as heavier than they actually are on Monday through Wednesday, but lighter than they actually are on Thursday through Friday. A bathroom scale might also be biased: it might show people as heavier than they actually are every day of the week. A bathroom scale might also be simultaneously noisy and biased: it might show people as heavier than they are every day of the week, but on Monday, show people as ten pounds heavier than they are, and on Tuesday, show people as five pounds heavier than they are.

It will be natural at this point to wonder about the relationship between bias and noise. In principle, the difference between systematic error (bias) and random error (noise) should not be obscure. But might biases help account for noise? The answer is emphatically yes.[13] Suppose, for example, that some doctors, in a hospital, show optimistic bias and thus fail to engage in sufficient testing. Suppose that other doctors, in the same hospital, show no such bias and thus order the right level of testing. An unshared bias might lead to noise within the hospital. In fact, whenever we observe noise at the system level, the reason might be an unshared bias. But for present purposes, I mean to emphasize the sharp difference between bias, in the form of systematic error, and noise, in the form of unwanted variability. In the administrative state, bias and noise can be serious problems, though we need far more research on both of them.

Markus K. Brunnermeier, Filippos Papakonstantinou & Jonathan A. Parker, *An Economic Model of the Planning Fallacy* (Nat'l Bureau of Econ. Rsch., Working Paper No. 14228, 2008) (exploring the planning fallacy in both theory and practice).

[9] *See generally* Ted O'Donoghue & Matthew Rabin, *Present Bias: Lessons Learned and To Be Learned*, 105 AM. ECON. REV. 273 (2015) (describing lessons learned through the study of present bias and the open questions that remain).

[10] *See* MAHZARIN R. BANAJI & ANTHONY G. GREENWALD, BLINDSPOT: HIDDEN BIASES OF GOOD PEOPLE xii (2013).

[11] *See* KAHNEMAN ET AL., *supra* note 2, at 6–7.

[12] *See id.* at 3–4.

[13] *See id.* at 69–93.

Because of my focus on bias and noise, I am more optimistic about the potential use of algorithms than are many people, including researchers focusing on their role in the administrative state.[14] Casual empiricism suggests that optimism about that role does not typically produce bright smiles and enthusiastic applause, while pessimism and serious warnings produce appreciative and knowing nods of assent. In light of that apparent fact, and acknowledging the force of many of the releveant concerns, I am acutely aware that I will, in a sense, be swimming against the current.

The remainder of this Article is organized as follows. Part I explores bias and noise. With respect to the former, I shall be focusing here on cognitive biases. It is an artifact of the English language, and a most unfortunate one, that the word "biases" includes both cognitive biases (my emphasis in Part I) and biases that involve illicit discrimination. In speaking of algorithms, we thus might be better off with different words—as in, c-biases and d-biases, respectively. C-biases, the subject of a massive literature in psychology and economics, refer to pervasive errors in judgment and decision-making[15]—as in, for example, optimism bias[16] and the planning fallacy.[17] Part I addresses c-biases, not d-biases, and shows that c-biases and noise both contribute to errors, though in different ways.

Part II turns to the silence of algorithms, noting that they eliminate noise and exploring in exactly what sense that is a good thing. (A preview: it is a better thing than one might think; noise is like a killer in a murder mystery whom one never notices until it is too late.) Part III explores the relationship between algorithms and biases, claiming that the statistical fallacies to which human beings are prone are likely to be avoided by algorithms, but that algorithms might make their own kinds of (cognitive) errors because of how they are designed. Part IV turns to the exceedingly complex questions raised by d-biases, distinguishing among disparate treatment, disparate impact, and racial balance. It aims to show exactly how, and in what sense, algorithms might encode discrimination. At the same time, it urges that this risk,

[14] *See, e.g.*, Megan Garcia, *Racist in the Machine: The Disturbing Implications of Algorithmic Bias*, WORLD POL'Y J., Winter 2016/2017, at 111, 112 (emphasizing the size and nature of algorithmic bias).

[15] *See generally, e.g.*, KAHNEMAN, *supra* note 8 (exploring noise in many contexts); RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: THE FINAL EDITION 23–88 (Penguin Books 2021) (2008) (discussing various biases).

[16] *See* SHAROT, *supra* note 7.

[17] *See supra* note 8.

sometimes realized in practice, should be taken as a reason for better algorithms, not for no algorithms—especially, perhaps, if our goal is to eliminate discrimination and injustice.

## I. NOISE, BIAS, AND THE ADMINISTRATIVE STATE

The human mind can be seen as a kind of scale or measuring instrument; it might turn out to be biased, noisy, or both.[18] To the extent that they are run by human beings, public institutions, including administrative agencies, are subject to c-biases.[19] They are also noisy.[20]

To be sure, it is necessary to be quite careful here. As a general rule, agencies do not base their judgments on intuitions. Agencies are highly likely to have processes and safeguards in place to reduce cognitive errors or to limit the effect of biases. For example, cost-benefit analysis can have precisely that consequence; one of its main goals is to discipline intuitions and to ensure that an assessment of consequences is the foundation of regulatory choices.[21] Cost-benefit analysis might reduce noise and bias at the same time—as, for example, in the case of a uniform value of a statistical life (a value that represents the monetary value that federal agencies assign to a human life).[22] Agencies might also, and often do, rely on rules or guidelines to reduce bias and noise.[23] The magnitude of both bias and noise will vary across agencies and functions and will depend in part on the nature and extent of the relevant processes and safeguards. The only point is that some bias, and some noise, are highly likely in the operations of the

[18] KAHNEMAN ET AL., *supra* note 2, at 3–6, 39–42.

[19] For a detailed discussion of the relationship between availability bias and risk regulation, see generally Timur Kuran & Cass R. Sunstein, *Availability Cascades and Risk Regulation*, 51 STAN. L. REV. 683 (1999).

[20] *See* Daniel Chen, Tobias J. Moskowitz & Kelly Shue, *Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires* 1–2 (Nat'l Bureau of Econ. Rsch., Working Paper No. 22026, 2016); KAHNEMAN ET AL., *supra* note 2, at 6–7.

[21] *See generally* Cass R. Sunstein, *Cognition and Cost-Benefit Analysis*, 29 J. LEGAL STUD. 1059 (2000) (urging that cost-benefit analysis can correct for c-biases).

[22] *See* Cass R. Sunstein, *The Value of a Statistical Life: Some Clarifications and Puzzles*, 4 J. BENEFIT-COST ANALYSIS 237, 237–41 (2013) (discussing the use of the value of statistical life and its foundations).

[23] For a classic study, see generally JERRY L. MASHAW, BUREAUCRATIC JUSTICE (1983), which emphasizes the role and value of rules in administrative adjudication.

administrative state, which may turn out to be extremely troubling. Indeed, we have evidence to suggest that they are exactly that.[24]

I now turn to the role of bias and noise within the executive branch. While the discussion will be mostly conceptual rather than empirical, my main goal is to demonstrate that bias and noise play a role in all human judgments, including those made by administrators. An understanding of that role will help pave the way toward an understanding of the promise of algorithms.

### *A. Cognitive Bias in the Administrative State*

Let us turn to administrative adjudicators and imagine that they are making some kind of judgment—say, about whether applicants for asylum or refugee status face "a well-founded fear of persecution on account of race, religion, nationality, membership in a particular social group, or political opinion."[25] Adjudicators might turn out to be biased in some general way. Perhaps because of a c-bias, such as availability bias, they might show consistent (and excessive) receptivity toward all applicants relative to the best understanding of the legal standard. Or perhaps because of availability bias, they might show consistent antipathy toward all applicants in a way that leads to a systematic bias relative to the best understanding of the legal standard. Of course, any such bias might be based in a value of some kind (such as skepticism about granting asylum), rather than a c-bias; but in any case, it might be counted as a bias.

More interestingly, the bias of an adjudicator might be *selective*. It might take the form of some kind of prejudice against, or in favor of, specific types of claimants—for example, those seeking asylum because of their religious affiliation or because of their political views.[26] Such a

[24] *See, e.g.*, Jaya Ramji-Nogales, Andrew I. Schoenholtz & Philip G. Schrag, *Refugee Roulette: Disparities in Asylum Adjudication*, 60 STAN. L. REV. 295, 301–02 (2007) [hereinafter Ramji-Nogales et al., *Refugee Roulette*]; Alafair S. Burke, *Improving Prosecutorial Decision Making: Some Lessons of Cognitive Science*, 47 WM. & MARY L. REV. 1587, 1590–92 (2006); Sjoerd Stolwijk & Barbara Vis, *Politicians, the Representativeness Heuristic and Decision-Making Biases*, 43 POL. BEHAV. 1411, 1427–29 (2020).

[25] 8 U.S.C. § 1101(a)(42).

[26] For evidence to this effect in the federal courts, see Kenny Mok & Eric A. Posner, Constitutional Challenges to Public Health Orders in Federal Courts During the COVID-19 Pandemic 3–4 (Aug. 1, 2021) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3897441 [https://perma.cc/GU63-24WK]. See generally Fatma E. Marouf, *Implicit Bias and Immigration Courts*, 45 NEW ENG. L. REV. 417 (2011), showing how implicit bias, with few safeguards to prevent it, unduly influences immigration decisions.

prejudice might be rooted in the availability heuristic,[27] the affect heuristic, or in some other rule of thumb. An adjudicator who is selective in her bias might produce consistent and excessive prejudice *against* some applicants, and consistent and excessive prejudice *in favor* of other applicants.

In the adjudicative system as a whole, we might find a general bias in the form of a systematic tendency toward excessive stringency or excessive leniency—or instead some kind of selective bias, as when certain applicants are treated with excessive stringency, and others with excessive leniency. Of course, administrative prosecutors might be biased too, and their biases might lead to a general or selective bias in the exercise of enforcement discretion. Prosecutors might, for example, target certain polluters for suspected Clean Air Act violations or certain employers for suspected Occupational Safety and Health Act violations in a way that reflects a general or specific bias.[28]

### *B. Noise in the Administrative State*

There is also a good chance that any system of administrative adjudication will be noisy in the sense that it will show unwanted variability in adjudicative or prosecutorial judgments.[29] Unwanted variability exists if identically situated applicants are treated differently merely because of the identity of the adjudicator or because the case comes before a particular adjudicator at time 1 rather than at time 2.[30] Compelling evidence of noise, so understood, can be found in the domain of refugee adjudications in particular; the system involves a kind of "[r]efugee [r]oulette" in which outcomes turn on the identity of

[27] *See infra* Part III.C.

[28] *Cf.* Dan P. Ly, *The Influence of the Availability Heuristic on Physicians in the Emergency Department*, 78 ANNALS EMERGENCY MED. 650, 650–53 (2021) (discussing how use of the availability heuristic by doctors leads some doctors to test more for conditions they have diagnosed recently compared to other doctors).

[29] As Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan explain,

> [T]he superior performance of the predicted judge suggests that, on net, the costs of inconsistency outweigh the gains from private information in our context. Whether these unobserved variables are internal states, such as mood, or specific features of the case that are salient and overweighted, such as the defendant's appearance, the net result is to create noise, not signal.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 242 (2018) [hereinafter Kleinberg et al., *Human Decisions*].

[30] *See* KAHNEMAN ET AL., *supra* note 2, at 43–54.

the particular person chosen to be the adjudicator.[31] As the authors of the leading article put it,

> [H]ow about a situation in which one judge is 1820% more likely to grant an application for important relief than another judge in the same courthouse? Or where one U.S. Court of Appeals is 1148% more likely to rule in favor of a petitioner than another U.S. Court of Appeals considering similar cases?

Welcome to the world of asylum law.[32]

These are exceptionally dramatic numbers, but in the world of administrative adjudication, noise is pervasive.[33] A particularly interesting source of noise is associated with the gambler's fallacy: after a series of approvals, administrative law judges are less likely to grant asylum than they are after a series of disapprovals![34] Prosecutors are likely to be noisy too.[35] In brief: *whenever there is human judgment in the administrative state, there is likely to be noise—and probably more than one might think*.[36]

To understand that proposition, we should distinguish among three kinds of noise. The first is "occasion noise," which exists if the same judge is influenced by self-evidently irrelevant features of the particular situation.[37] Occasion noise is *intrapersonal*.[38] Suppose, for example, that an administrative adjudicator decides differently on Monday than on Friday, or in the morning than in the late afternoon, or after a victory than after a loss by the local football team, or on a warm than on a cold day. Occasion noise has been found in startling

[31] *See* Ramji-Nogales et al., *Refugee Roulette*, *supra* note 24, at 295, 301–02.

[32] *Id.* at 301. Refugee roulette can be found many places. *See, e.g.*, Andrew Burridge & Nick Gill, *Conveyor-Belt Justice: Precarity, Access to Justice, and Uneven Geographies of Legal Aid in UK Asylum Appeals*, 49 ANTIPODE 23, 23–30 (2017) (describing how the U.K. asylum appeal success rate is affected by the location of the asylum seeker and corresponding access to legal representation).

[33] On the general phenomenon and potential correctives, see generally MASHAW, *supra* note 23.

[34] *See, e.g.*, Chen et al., *supra* note 20, at 1–3.

[35] *See, e.g.*, David M. Uhlmann, *Prosecutorial Discretion and Environmental Crime*, 38 HARV. ENV'T L. REV. 159, 164 (2014) (discussing how prosecutors exercise discretion in choosing which environmental crimes to prosecute). *See generally* ANGELA J. DAVIS, ARBITRARY JUSTICE: THE POWER OF THE AMERICAN PROSECUTOR (2007) (discussing how prosecutorial discretion, without sufficient public scrutiny and oversight to ensure fairness, has led to wide disparities in how prosecutors treat different cases).

[36] This is an adaptation of the central theme of KAHNEMAN ET AL., *supra* note 2.

[37] *Id.* at 366–67.

[38] *See id.* at 367.

places.[39] In many contexts, we can be confident that occasion noise exists, though we cannot be confident about its magnitude.[40] For the administrative state, there is a large research project here.

The second kind of noise is "level noise," which is interpersonal.[41] If some judges, entrusted with making decisions about asylum, are more severe than others, and systematically so, we will have level noise at the system level. Similarly situated people will be treated differently, because of a kind of lottery. "Pattern noise," the third kind of noise, is also interpersonal, but it is very different from and more subtle than level noise.[42] Level noise comes from a systematic tendency for some people to be more severe than others; by contrast, pattern noise comes not from any such tendency, but from different patterns of severity and leniency.[43] Suppose that judges A and B are receptive to asylum on religious grounds, but not receptive to asylum on nationality grounds, and that judges C and D show the opposite pattern. The system will show significant noise. But the reason is not a *general* difference in the level of severity as between judges A and B on the one hand and judges C and D on the other. The reason is a difference in their respective patterns of severity and leniency. It would be exceedingly valuable to know the magnitude of occasion noise, level noise, and pattern noise in administrative processes, and especially in administrative adjudication, but there is little question that all three forms of noise are pervasive.

We are now in a position to understand the concepts of bias and noise in judgment and to see why they are almost certainly playing a major role in the administrative state. How might algorithms help?

[39] *See, e.g.*, *id.*

[40] For evidence that it might well be significant, see Chen et al., *supra* note 20, at 1–2, finding that, in asylum cases, up to "two percent of decisions [are] reversed purely due to the sequencing of past decisions, all else equal."

[41] *See* KAHNEMAN ET AL., *supra* note 2, at 73–74 (discussing how judges impose sentences with different levels of severity, which may be based on factors such as their opinions about the goals of sentencing, their geographic locations, and their political ideologies).

[42] *Id.* at 74–76 (discussing how judges may be harsher or more lenient than they usually are when sentencing in particular cases).

[43] *Id.* at 73–76.

## II. ALGORITHMS ARE SILENT

Algorithms are not noisy. By their very nature, they are silent.[44] If an applicant seeks asylum, the algorithm will offer the same answer whether it is Monday or Wednesday or January or June. Someone whose asylum application follows five successful applications will not be treated differently from someone whose application follows five unsuccessful applications. There is no occasion noise because the occasion cannot, and does not, matter. And because the level is the same across applications, there is no level noise. For the same reason, algorithms cannot, and will not, display pattern noise. An algorithm with identical source code will not produce a different result in identical cases.

It might be tempting to think that these points are not terribly important, and that the silence of algorithms is not much to celebrate. One reason involves the apparent lessons of intuition; another involves the concern about bias. As an explanation of administrative judgments, or of failure in the administrative state, bias has a kind of charisma. It is like the singer who commands the stage. One more time: Noise, by contrast, is like the character in a movie who seems boring and trivial—but who turns out to be the killer.

It might be tempting to think that across a system, noise cancels out. If judges in an adjudicative system are too stringent half of the time and too lenient half of the time, there is no total bias—and perhaps things are not so terrible. But in fact, things are very terrible. To know the total error, you must add the errors on both sides. In a noisy system of administrative adjudication, that might be a large number. If one thousand people are getting asylum but do not deserve it, and one thousand people are not getting asylum but do deserve it, we have a serious problem. A noisy, unbiased administrative agency might produce a higher level of total error than a biased, quiet administrative agency. To know, we need to find out how noisy the noisy agency is and how biased the biased agency is.[45]

[44] *See generally* Jens Ludwig & Sendhil Mullainathan, *Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System*, 34 J. ECON. PERSPS. 71 (2021) (exploring different kinds of bias and methods for correcting algorithmic bias).

[45] As I have noted, an unshared bias can be responsible for noise, but let us bracket that possibility here. *See supra* Introduction. The basic point is an agency that shows a small systematic bias might show less total error than does an agency that shows a lot of random variability (without *any* systematic bias).

It is worth pausing over these points. As we have seen, a quiet algorithm will ensure equal treatment; it will prevent a Kafkaesque situation in which outcomes depend on a lottery, in which the most important moment might be the hour in which a particular person is chosen as an adjudicator or a prosecutor. A quiet algorithm will not fall prey to the gambler's fallacy. That is a large gain. And a quiet algorithm will also reduce mistakes. Even if it is biased, a scale that is not noisy—one that, say, *always* shows people as one pound heavier than they actually are—will produce far less in the way of total error than if it is noisy as well.

There are important qualifications. A noise-free system might be nothing to celebrate. Consider the question of mercy, understood as a departure from the rules, driven by an understanding of the circumstances of the individual case. An algorithm might be merciless. Whether we are speaking of sentencing, asylum, licenses, permits, or health care benefits,  a noise-free algorithm might be intolerably rigid, in a way that ensures that it will not take account of particular circumstances, or listen to people who are making some kind of plea. This is an important concern, one that calls up the interest in individual dignity.

To handle that concern, a detailed discussion would be necessary. For now, consider two points. First, algorithms might be designed so as to be merciful, in an important sense. They could be programmed to be highly attuned to particular circumstances, certainly if they are making predictive judgments, and also if they are making judgments about appropriate punishment. This form of attunement may or may not be superior to the attunement of human beings. Second, mercy might produce unacceptable inequality and also error. If (some) people are being given light punishments for terrible crimes, or being granted asylum when they do not deserve it, or licenses for which they lack appropriate consequences, the systemic consequences might be serious and unwanted. The capacity for mercy, and for individualized treatment, are relevant considerations, but with respect to the use of algorithms, they should not be taken as trump cards.

With respect to the charisma of bias, it should be obvious that an algorithm that is very biased, but not at all noisy, might well produce an unacceptably large number of errors, even if there is no problem of unequal treatment. Suppose, for example, that a scale consistently shows people as ten pounds heavier than they actually are. Or suppose that an algorithm is unrealistically optimistic: it consistently predicts that tasks will take 20 percent less time than they actually do. (Note that if it does so, it is because of how human beings have programmed it.) Or suppose that the algorithm is unrealistically pessimistic: it greatly overstates the likelihood that certain taxpayers are cheating. (Same proviso.) Or suppose that an algorithm understates the adverse effects of certain disabilities, such as depression, so that it wrongly concludes that certain kinds of people, with specific characteristics, are not entitled to disability benefits. (One more time.) The elimination of noise is both important and good, but it is no guarantee of accuracy.

A number of years ago, I encountered a chess program on an international flight. I am not a very good chess player, so I chose the program's easiest level. I quickly learned that the program's algorithm ensured that it would always seek to put the other player in check, regardless of whether that was a good idea. Call it the "Put the Other Player in Check Heuristic," leading to the "Put the Other Player in Check Bias." The algorithm was not noisy. It *always* used that heuristic and *always* displayed that bias. Because it did so, it was easy to defeat, even for this far-from-good chess player. (The airline apparently wanted happy travelers.) Noiselessness can be a large virtue, but it is hardly a cure-all—which brings us to the next topic.

## III. ALGORITHMS AND BIAS

Are algorithms biased? In what respect? And why? These are large questions, and there are no simple answers. My principal claim here is that algorithms can overcome the harmful effects of c-biases. These biases sometimes have a strong hold on people whose job it is to avoid them, and whose training and experience might be expected to equip them to do so.[46] In particular, the administrative state is presented with a large number of prediction problems, for which c-biases can lead people astray. Algorithms can be a great help, and insofar as we are speaking of the standard c-biases, they might even be a complete corrective.[47] At the same time, it is emphatically true that

[46] *See* POHL, *supra* note 3, at 4–5.

[47] For valuable discussions on how algorithmic predictions help to understand and reduce physicians' over- and underuse of testing in the medical field, see generally Sendhil Mullainathan & Ziad Obermeyer, *Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care* 4–5 (Nat'l Bureau of Econ. Rsch., Working Paper No. 26168, 2021); David Arnold, Will S. Dobbie & Peter Hull, *Measuring Racial Discrimination in Algorithms* 2 (Nat'l Bureau of Econ. Rsch., Working Paper No. 28222, 2021); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Ziad Obermeyer, *Prediction Policy Problems*, 105 AM. ECON. REV. 491, 491 (2015). On availability bias in medicine, see Ping Li, Zi yan Cheng & Gui lin Liu, *Availability Bias Causes*

algorithms can go wrong, and if they are designed in certain ways, they can encode c-biases of their own—recall the Put the Other Player in Check Bias.

### A. *Statistical Predictions, Cognitive Biases, and Discrimination*

Some of the oldest and most influential work in behavioral science demonstrates that statistical prediction often outperforms clinical prediction. One reason for this involves c-biases on the part of clinicians, which taint their predictions.[48] Algorithms can be seen as a modern form of statistical prediction, so if they avoid biases, no one should be amazed. The central point is that algorithms should not fall prey to the kinds of c-biases to which human beings are prone, because they rely on statistical predictors. Unless they are programmed to do so, they will not use the availability heuristic; they will not be susceptible to anchoring; and they will not be unrealistically optimistic. What I hope to add here is a concrete demonstration of these points in an important context, with some general remarks designed to address—not at all to dismiss—the concern that algorithms are "biased."

It is true, of course, that algorithms might go wrong if they are built in such a way as to encode c-biases, or if they get the wrong data inputs. Suppose that an algorithm is asked to find the cheapest hotel in some area. If the algorithm has been given data only about Boston, people who are willing to travel outside of Boston will not receive the information they need. And if the algorithm has been given information only about the Four Seasons and the Ritz, the algorithm will not be particularly helpful to those who want to look elsewhere. (We can make similar points about the kinds of judgments made by administrative agencies. If agencies are receiving the wrong inputs, their judgments may be biased, and if the inputs reflect availability bias or optimism bias, their judgments will misfire.) The only point is that unless they are programmed to do so, algorithms will not make statistical errors—which is not, to be sure, a guarantee that they will not err at all.

*Misdiagnoses by Physicians: Direct Evidence from a Randomized Controlled Trial*, 59 INTERNAL MED. 3141, 3141 (2020), which found a significant role for availability bias among doctors.

[48] *See generally* PAUL E. MEEHL, CLINICAL VERSUS STATISTICAL PREDICTION: A THEORETICAL ANALYSIS AND A REVIEW OF THE EVIDENCE (2013) (comparing clinical prediction to statistical prediction and finding that the latter is usually better).

It is correct to emphasize the importance of the "unless they are programmed to do so" proviso; algorithms can be programmed to use heuristics and to make errors. It is also entirely right to object that algorithms can encode d-bias and hence perpetuate discrimination, perhaps because they use discriminatory inputs, perhaps because they predict something that is infected by discrimination.[49] A discriminatory input, for example, may be arrest records in a place in which people of color are more likely to be arrested than white people are because of biased policing. Or suppose that employers are more likely to fire women than men after a year of employment; suppose too that the difference is a product of discrimination. If so, an algorithm that predicts who will be retained after a year will favor men.

Nonetheless, I suggest, algorithms can be designed by human beings so as to avoid d-biases—racial or other discrimination in its most unambiguously unlawful forms. I also suggest that the topic of algorithmic bias raises exceedingly hard questions about how to understand the very idea of discrimination and to balance competing social values.[50]

Complaints about algorithmic bias often focus on race and sex discrimination.[51] The word "discrimination" can of course be understood in many different ways.[52] When we find algorithmic bias, or something close to it, the reason lies in emphatically human decisions, not in artificial intelligence as such.[53] For that reason, it might turn out to be relatively simple to ensure that algorithms do not discriminate in the way that U.S. law most squarely and least controversially addresses.[54] It is less simple to deal with outcomes that

[49] *See, e.g.*, Arnold et al., *supra* note 47.

[50] The latter question is explored in Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113 (2018) [hereinafter Kleinberg et al., *Discrimination in the Age of Algorithms*] (urging that algorithms can be more transparent than human beings and thus serve to reduce discrimination).

[51] *See, e.g.*, Noel Sharkey, *The Impact of Gender and Race Bias in AI*, ICRC HUMANITARIAN L. & POL'Y BLOG (Aug. 28, 2018), https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai [https://perma.cc/6RL9-KVFN]; Alex Engler, *Auditing Employment Algorithms for Discrimination*, BROOKINGS (Mar. 12, 2021), https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination [https://perma.cc/M27G-TYR6].

[52] *See generally* David A. Strauss, *Discriminatory Intent and the Taming of* Brown, 56 U. CHI. L. REV. 935 (1989) (exploring different possible meanings of discrimination and discriminatory intent).

[53] *See* Ludwig & Mullainathan, *supra* note 44, at 82–88.

[54] *Cf.* Washington v. Davis, 426 U.S. 229, 239 (1976) (ruling that discriminatory intent, not simply disparate impact, is necessary for an equal protection violation).

concern many people, including disparate impact and an absence of racial balance as such. As we shall see, algorithms allow new transparency about some difficult tradeoffs.[55]

### *B. Bail, Flight Risk, and Crime*

The principal research on which I will focus, enlisted here as a kind of proof of concept, comes from Professors Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, who explore judges' decisions on whether to release criminal defendants pending trial.[56] 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.[57] This may or may not be characterized as, strictly speaking, a question about the administrative state, but it is analytically close to many issues that administrative prosecutors and adjudicators have to answer, at least insofar as they are dealing with prediction problems.

It should be obvious that the decision whether to release defendants has large consequences. If defendants are incarcerated, the long-term consequences can be severe. Their lives can be ruined. But if defendants are released, they might flee the jurisdiction or commit crimes.

In some states, the decision whether to allow pretrial release turns on a single factor: flight risk.[58] To answer that question, 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,[59] 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 is likely accurately to predict the second as well.[60])

[55] *See* Kleinberg et al., *Discrimination in the Age of Algorithms*, *supra* note 50, at 119–20.

[56] Kleinberg et al., *Human Decisions*, *supra* note 29, at 239.

[57] *See id.*

[58] *See* Lauryn P. Gouldin, *Defining Flight Risk*, 85 U. CHI. L. REV. 677, 696 (2018); *see also id.* at 682 (defining flight risk as "the risk that a defendant will fail to appear for a future court date").

[59] *See* Kleinberg et al., *Human Decisions*, *supra* note 29, at 239.

[60] *See id.* at 273–75; *see also* John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725, 1733–34 (2018) (discussing the history of bail and that it was controversially used as a way to prevent people from committing further crimes in addition to its accepted flight risk rationale). *But see* Lauryn P. Gouldin, *Disentangling Flight Risk from Dangerousness*, 2016 BYU L. REV. 837, 843 (making

Kleinberg and his colleagues built an algorithm that uses as inputs the same data available to judges at the time of the bail hearing, such as prior criminal history and current offense.[61] Their central finding is that along every dimension that matters, the algorithm does much better than real-world judges.[62] Among other findings:

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.[63] Alternatively, use of the algorithm could ensure that there is no increase in crime while also reducing detention rates by as much as 41.9 percent.[64] 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. 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. Human judges make a major mistake by releasing many people identified by the algorithm as especially high risk or 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.[65] Those defendants fail to reappear in court 56.3 percent of the time.[66] They are rearrested at a rate of 62.7 percent.[67] Judges thus 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 the riskiest defendants. If it were, the strictest judges could jail as many people as they now do, but with a 75.8 percent reduction in crime.[68] Alternatively, they could

"constitutional, statutory, and policy-based arguments to illustrate why . . . disentangling [flight risk from dangerousness] is integral to [bail] reform efforts").

[61] *See* Kleinberg et al., *Human Decisions*, *supra* note 29, at 239.

[62] *See id.* at 241–42, 284–86.

[63] *Id.* at 241.

[64] *Id.*

[65] *Id.* at 240.

[66] *Id.*

[67] *Id.*

[68] *Id.*

> keep the current crime reduction and jail only 48.2 percent as many people as they now do.[69]

A full account of why the algorithm outperforms judges would require an elaborate treatment. But for my purposes here, one part of the explanation is particularly revealing. As the second point above suggests, many judges, not merely those that are most strict, do poorly with the highest-risk cases.[70] The reason is an identifiable bias; call it "Current Offense Bias." The bias comes in turn from an identifiable heuristic; call it the "Current Offense Heuristic." On this count, Kleinberg and his colleagues restrict their analysis to two brief sentences, but those sentences have immense importance:

> We find that judges struggle not so much with the middle of the distribution, but instead with one tail: the highest-risk cases. . . . That is, judges treat many of these high-risk cases as if they are low risk. We have also examined the characteristics that define these tails. Judges are most likely to release high-risk people if their current charge is minor, such as a misdemeanor, and are more likely to detain low-risk people if their current charge is more serious. Put differently judges seem to be (among other things) overweighting the importance of the current charge.[71]

As it turns out, then, human judges make two fundamental mistakes. First, they treat high-risk defendants as if they are low risk when the current charge is *relatively minor*—for example, it may be a misdemeanor. Second, they may treat low-risk defendants as if they are high risk when the current charge is *especially serious*—for example, it may be a felony. 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 the current charge. The fact that judges release a number of the high-risk defendants is attributable, in large part, to their overweighting the current charge when it is not especially serious. The general point should not be obscure: algorithms outperform human judges in this context, and the limitations of human judges, rooted partly in a cognitive error, produce terrible consequences.

[69] *Id.*

[70] *Id.* at 284.

[71] *Id.* (citations omitted).

### *C. Availability Bias and Its Cousins*

Current Offense Bias is of general, rather than particular, interest. It shows that when human beings suffer from a c-bias, a well-designed algorithm attempting to solve a prediction problem can do much better.[72] It is worth emphasizing that we are dealing with trained and experienced people, not novices. They are experts. Nonetheless, they suffer from c-biases that produce severe and systematic errors. For example, closely related research shows that with respect to heart disease, an algorithm greatly outperforms human physicians: doctors test many patients when they should not do so, and they do not test many patients when they should do so, leading to both excessive costs and adverse health events.[73] One key reason appears to be that doctors overweight salient information (immediate symptoms and demographics, as compared to past laboratory studies and vital signs).[74] Another key reason is that doctors rely on something close to the representativeness heuristic, overweighting symptoms that seem representative of a heart attack, such as chest pain and shortness of breath.[75] These are remarkably similar findings to those in the bail study.

For purposes of thinking about the role of algorithms in the administrative state, judges' use of the Current Offense Heuristic, and the algorithm's different approach, have broader interest still. To be sure, it would be valuable to understand what, precisely, lies behind Current Offense Bias; it might well be closely related to the affect heuristic, in the sense that the current offense might well produce an affective reaction that operates as a shortcut for judgments about flight risk. But on its own terms, Current Offense Bias is plausibly understood as a close cousin of availability bias: individual judgments about probability are frequently based on whether relevant examples are easily brought to mind.[76] Both biases involve *attribute substitution*.[77]

[72] *See* Mullainathan & Obermeyer, *supra* note 49, at 4, 22, 38–39 (noting algorithms can help correct both over- and undertesting for blockages that can lead to heart attacks); Kleinberg et al., *Human Decisions*, *supra* note 29, at 240–42.

[73] Mullainathan & Obermeyer, *supra* note 47, at 4–5.

[74] *Id.* at 5, 34.

[75] *See id.* at 4–5, 32–33.

[76] *See* Amos Tversky & Daniel Kahneman, *Availability: A Heuristic for Judging Frequency and Probability*, *in* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES, *supra* note 4, at 163, 163 [hereinafter Tversky & Kahneman, *Availability*] (describing the availability bias).

[77] *See* Daniel Kahneman & Shane Frederick, *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, *in* HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE

Availability bias is a product of the availability heuristic, which people use to solve prediction problems.[78] They substitute a relatively easy question ("Does an example come to mind?") for a difficult one ("What is the statistical fact?"). The Current Offense Heuristic poses a relatively easy question. Algorithms will not substitute an easy question for a hard question, at least not in the sense that human beings do; they will ask the hard question.

Because of the availability heuristic, many people are likely to think that more words on a random page end with the letters "ing" than have "n" as their penultimate letter[79]—even though a moment's reflection will show that this could not possibly be the case. An algorithm would not make this mistake. Moreover, "a class whose instances are easily retrieved will appear more numerous than a class of equal frequency whose instances are less retrievable."[80] Consider a simple study involving a list of well-known men and women, with the same number of each gender, that asked participants whether the list contains more names of women or more names of men.[81] In lists in which the men were especially famous, participants thought that there were more names of men, whereas in lists in which the women were more famous, participants thought that there were more names of women.[82]

This is a point about how *familiarity* can affect the availability of instances and thus produce mistaken judgments (including 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."[83] *Recency* also

JUDGMENT 49, 53 (Thomas Gilovich, Dale Griffin & Daniel Kahneman eds., 2002) (describing attribute substitution as "when an individual assesses a specified target attribute of a judgment object by substituting another property of that object—the heuristic attribute—which comes more readily to mind" (emphasis omitted)); *see also* KAHNEMAN, *supra* note 8 (distinguishing between rapid, intuitive thinking and deliberative thinking).

[78] *See generally* Tversky & Kahneman, *Availability*, *supra* note 76 (outlining the availability heuristic and evidence on its behalf).

[79] *See id.* at 166–68.

[80] Tversky & Kahneman, *Judgment Under Uncertainty*, *supra* note 4, at 11.

[81] *Id.*

[82] *Id. See generally* Drew Fudenberg & David K. Levine, *Learning with Recency Bias*, 111 PROC. NAT'L ACAD. SCIS. 10826 (2014) (demonstrating the validity of recency bias).

[83] Tversky & Kahneman, *Judgment Under Uncertainty*, *supra* note 4, at 11.

matters. Because recent events tend to be more easily recalled, they will have a disproportionate effect on probability judgments.[84] Availability bias thus helps account for "recency bias."[85]

Current Offense Bias can be understood as a sibling to recency bias. The current offense is, of course, the most recent one, which means that Current Offense Bias might actually be a form of recency bias. In addition, the current offense will be highly salient, which means that it could loom especially large in judicial judgment. It might or might not be right to deem it "familiar," but the current offense will, of course, attract the judge's attention; it is, after all, the offense for which the defendant has been arrested. For all these reasons, it might have (and apparently does have) an outsized effect on how judges proceed.

In many domains, people must solve prediction problems; availability bias in those domains can lead to damaging and costly mistakes. Whether people will anticipate future natural disasters is greatly affected by recent experiences.[86] In the aftermath of an earthquake, purchases of earthquake insurance rises sharply; they decline steadily from that point as vivid memories recede.[87] Note that the use of the availability heuristic in these contexts is hardly irrational. Both insurance and precautionary measures can be expensive. What has happened before often seems to be the best available guide to what will happen again. The problem is that the availability heuristic can lead to both excessive fear and neglect. And the point is not limited to ordinary people seeking answers to hard questions; the availability heuristic can affect administrators as well, in part because they are human, and in part because they are subject to democratic checks.[88] The last point is worth underlining. If the public is greatly concerned about some issue, perhaps because of the availability heuristic, it might

[84] Fudenberg & Levine, *supra* note 82, at 10826.

[85] *See* Robert H. Ashton & Jane Kennedy, *Eliminating Recency with Self-Review: The Case of Auditors' 'Going Concern' Judgments*, 15 J. BEHAV. DECISIONMAKING 221, 222 (2002) (describing how recency bias's impacts can be compounded by limited access to information).

[86] *See* PAUL SLOVIC, THE PERCEPTION OF RISK 40 (Ragnar E. Löfstedt ed., 2000).

[87] *See, e.g.*, Howard Kunreuther, *The Role of Insurance in Reducing Losses from Extreme Events: The Need for Public–Private Partnerships*, 40 GENEVA PAPERS 741, 745 (2015) (discussing earthquake insurance coverage in California after the 1994 Northridge earthquake).

[88] *See generally* Kuran & Sunstein, *supra* note 19 (analyzing availability cascades, "collective belief formation [processes] by which an expressed perception triggers a chain reaction that gives the perception increasing plausibility through its rising availability in public discourse," and suggesting reforms to address their hazards, "includ[ing] new governmental structures designed to [insulate] civil servants" from these pressures).

demand an immediate response, even if the underlying risk is low.[89] If the public is not exercised about some issue, perhaps because of the availability heuristic, it might not demand an immediate response, even if the underlying risk is high.[90]

If the goal is to make accurate factual judgments, the use of algorithms can be a great boon. For both private and public institutions, algorithms can eliminate the effects of c-biases. Suppose the question is whether to hire a job applicant; whether a project will be completed within six months; whether a taxpayer is likely to have cheated; whether a particular individual has a well-founded fear of persecution; whether a community faces a flood risk of a certain magnitude, or a risk of fire. In all of these cases, some kind of c-bias may distort human decisions, including those of administrators. It is possible 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. On this count, algorithms have extraordinary promise. In addition to eliminating noise, they can reduce the effects of c-biases and thus save both money and lives. Recall the central point: algorithms will not fall prey to statistical biases, and they will not use the cognitive heuristics used by human beings. These heuristics generally work well, but they can lead to severe and systematic errors.

## IV. ALGORITHMS AND DISCRIMINATION

There is a great deal of concern that algorithms might discriminate or promote discrimination on illegitimate grounds, such as race or sex.[91] The concern appears to be growing, in part because of real evidence that algorithms can incorporate, and perpetuate, some kind of bias.[92]

[89] *See* Cass R. Sunstein, *The Availability Heuristic, Intuitive Cost-Benefit Analysis, and Climate Change*, 77 CLIMATE CHANGE 195, 196–97 (2006).

[90] *Id.*

[91] *See, e.g.*, Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1024–25 (2017); Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/6JT5-UQH9].

[92] *See, e.g.*, Arnold et al., *supra* note 47; *see also* Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used To Manage the Health of Populations*, 366 SCIENCE 447, 447 (2019) (describing how a widely used health system algorithm exhibits racial discrimination). A terrific, clarifying discussion can be found in Ludwig & Mullainathan, *supra* note 44, at 82–88.

"Algorithmic bias" has become a common term.[93] To understand the problem, we need to go behind the evidence to understand why, when, and in what sense algorithms are biased. (Whether human beings are less biased, or more so, is a fair and important question.) The possibility that algorithms will promote discrimination raises an assortment of difficult issues, on which the bail research casts some new light.[94] Above all, the research suggests a powerful and simple point: use of algorithms will reveal, with great clarity, the need to make tradeoffs between the value of equality and other important values, such as public safety.

### *A. A Little Law, Very Briefly*

U.S. discrimination law has long been focused on two different problems. The first is disparate treatment; the second is disparate impact.[95] The Constitution, and essentially all civil rights laws, forbid disparate treatment.[96] The Constitution does not forbid practices that have a disparate impact,[97] but some civil rights statutes do.[98] I will be painting with a broad brush here, with the goal of setting out foundational principles for assessing whether and when algorithms might be said to be discriminatory as a matter of law.

1. *Disparate Treatment.* The Constitution forbids disparate treatment along a variety of specified grounds, above all race and sex.[99] The prohibition on disparate treatment reflects a commitment to a kind of neutrality. When such a prohibition is in place, public officials are not permitted to favor members of one group over another—unless, perhaps, there is a strong and sufficiently neutral reason for doing so, demonstrating that there is no favoritism at all.

[93] *See, e.g.*, *Algorithmic Bias Initiative*, CHI. BOOTH: CTR. FOR APPLIED A.I., https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias [https://perma.cc/GTX9-JSMX].

[94] *See* discussion *supra* Part III.B.

[95] For an overview, see Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 694 (2016).

[96] *See, e.g.*, Washington v. Davis, 426 U.S. 229, 239 (1976); Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 272 (1979).

[97] *See Washington*, 426 U.S. at 239.

[98] *See* Griggs v. Duke Power Co., 401 U.S. 424, 434–36 (1971) (interpreting Title VII of the 1964 Civil Rights Act).

[99] *See Feeney*, 442 U.S. at 272–73.

In extreme cases, the existence of disparate treatment is obvious because a facially discriminatory practice or rule can be shown to be in place—for example, a written policy stating "no women may apply." In other cases, no such practice or rule can be identified, and for that reason, violations are far 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 (for example, men) or to disfavor another (for example, women). To police discrimination, the legal system is required to use whatever tools it has to discern the motivation of decision-makers. Some of those tools might not be adequate to the problem at hand, even if discrimination in fact exists.

Violations of the prohibition on disparate treatment might arise because of explicit racial- or gender-based prejudice, sometimes described as "animus."[100] Explicit prejudice, on the part of human beings, might not be easy to uncover. Alternatively, such violations might arise because of unconscious prejudice, operating outside of the awareness of the decision-maker; unconscious prejudice is sometimes described as an "implicit bias."[101] 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 generally deplore. When an unconscious prejudice is at work, it might be especially difficult for the legal system to uncover it.

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—say, people of color or women—the requirement or practice must be adequately justified.[102] 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 women, they will be invalidated unless they can be shown to have a strong connection to the actual requirements of the

[100] *See generally* Susannah W. Pollvogt, *Unconstitutional Animus*, 81 FORDHAM L. REV. 887 (2012) (proposing a doctrinal definition of "animus" based on existing case law).

[101] *See* Samuel R. Bagenstos, *Implicit Bias, "Science," and Antidiscrimination Law*, 1 HARV. L. & POL'Y REV. 477, 477 (2007).

[102] *See Griggs*, 401 U.S. at 436; *Feeney*, 442 U.S. at 273.

job. Under some statutes, the defenders of such practices must show that the practices are justified by "business necessity."[103]

The theory behind disparate impact remains sharply disputed.[104] On one view, the goal is to ferret out disparate treatment. If an employer has adopted a practice with disproportionate adverse effects on women, 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. To make sense of this idea, we would need to ask something about the meaning of "discriminatory intent" in the relevant context. Under the Constitution, the Supreme Court has said that the question is whether the relevant decision was made "'because of,' not merely 'in spite of'" its discriminatory effect.[105] If a discriminatory effect is severe and very hard to justify in nondiscriminatory terms, perhaps we can infer that it was sought, thus satisfying the "because of" requirement.

Alternatively, we might understand the idea of discriminatory intent more broadly and ask a kind of "reversing the groups" question: *Would the decision have been made if, for example, the adverse effect was imposed on men rather than women?*[106] This question might be seen as a way of picking up on the problem of selective concern and indifference, which is arguably a form of discriminatory motive.[107] However we understand that kind of motive, a disparate impact test might be taken as a way of ferreting it out.

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.[108] Suppose, for example, that an agency's enforcement system disproportionately burdens people of color. If so, it might be thought necessary for those who adopt such practices to demonstrate that they have a good and sufficiently neutral

[103] 42 U.S.C. § 2000e-2(k)(1)(A)–(B).

[104] *See, e.g.*, Reva B. Siegel, *Foreword: Equality Divided*, 127 HARV. L. REV. 1, 2–4 (2013) (describing and critiquing the development of equal protection doctrine); Girardeau A. Spann, *Disparate Impact*, 98 GEO. L.J. 1133, 1135–37 (2010) (criticizing the Court's narrowing of the disparate impact doctrine); Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 706–07 (2006) (arguing that disparate impact theory is not correct).

[105] *See Feeney*, 442 U.S. at 279.

[106] *See* Strauss, *supra* note 52, at 956–57.

[107] *See* McCleskey v. Kemp, 481 U.S. 279, 336 (1987) (Brennan, J., dissenting).

[108] *See generally* Cass R. Sunstein, *The Anticaste Principle*, 92 MICH. L. REV. 2410 (1994) (suggesting that the Constitution's Equal Protection Clause might be understood as an attack on a caste system).

reason for doing so.[109] Because of its breadth, this understanding of the grounds for the disparate impact test is more contentious.

How do these points bear on the use of algorithms, and on the question of whether there is algorithmic bias? The answer is that we need to ask whether and when algorithms might cause either disparate treatment or disparate impact. That question might not have a simple answer.

### *B. Human Beings, Algorithms, and Discrimination*

1. *Human Judges*. In the context of bail decisions, we would have disparate treatment if it could be shown that judges discriminate against people of color, either through a formal practice or through a demonstrable discriminatory motive (established perhaps with some kind of extrinsic evidence).[110] On the other hand, we would have disparate impact if it could be shown that some factor, rule, or decision (taking account, for example, of employment history) had a disproportionate adverse effect on people of color; 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 clear proof of either, in the sense of existing law. For Blacks and Hispanics , the detention rate is 28.6 percent.[111] More specifically, Black defendants are detained at a rate of 31 percent, and Hispanic defendants are detained at a rate of 25 percent.[112] The detention rate for white individuals is between those two figures.[113]

2. *The Algorithm.* Importantly, the algorithm is, by design, blind to race. Whether a defendant is Black or Hispanic is not one of the factors that it considers in assessing flight risk. To that extent, we appear to have no problem of disparate treatment. We do not have explicit prejudice, and we do not have implicit prejudice; the algorithm is subject to neither. The point generalizes to the use of algorithms in

[109] Recall, however, that no such justification is necessary under the Constitution.

[110] *See* David Arnold, Will Dobbie & Crystal S. Yang, *Racial Bias in Bail Decisions*, 133 Q.J. ECON. 1885, 1886 (2018).

[111] *See* Kleinberg et al., *Human Decisions*, *supra* note 29, at 277.

[112] *Id.*

[113] *Id.*

administrative agencies, and that is an important gain. 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 directed to match the judges' overall detention rate, there might seem to be no obvious problem; its numbers, with respect to race, look quite close to the corresponding numbers for those judges.[114] Its overall detention rate for Blacks or Hispanics is 29 percent, with a 32 percent rate for Black defendants and 24 percent for Hispanic defendants.[115] At the same time, the crime rate drops, relative to judges, by a whopping 25 percent.[116] It would be fair to say that on any view, the algorithm is not a discriminator *if compared to human judges*. There appears to be no disparate treatment. It would be challenging to find disparate impact under existing principles. And in terms of outcomes, it is not worse along the dimension of racial disparities. (Whether the numbers are nonetheless objectionable is a fair and separate question.)

The authors 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 Blacks and Hispanics. Suppose that the algorithm is constrained so that the detention rate for Blacks 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[117]—significantly but not massively lower than the 25 percent rate achieved without the instruction to equalize the release rate. One finding is particularly revealing: if the algorithm is instructed to produce the same crime rate that judges currently achieve, it will jail 40.8 percent fewer Black defendants and 44.6 percent fewer Hispanic defendants.[118] It does this because it detains many fewer people, due to its focus on the riskiest

[114] *Id.*

[115] *Id.*

[116] *Id.*

[117] *Id.*

[118] *Id.* at 278.

defendants; many Blacks and Hispanics benefit from its more accurate judgments.[119]

The most important point here does 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. It might also show that racial discrimination, of one or another sort, is real.[120] And of course something similar could be said for other areas, outside of criminal justice, with which the administrative state regularly deals.

### *C. Larger Considerations*

If we say that algorithms correct for biases, we might be speaking of c-biases (such as Current Offense Bias). The case of d-biases is more challenging.[121] To be sure, disparate treatment should be preventable; algorithms do not have motivations, and they can be designed so as not to draw explicit lines on the basis of race or sex, or to take race or sex into account. (This is an ambiguous phrase, and I will return shortly to the ambiguity.) 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—say, flight risk, educational attainment, or job performance—but that factor might have a disparate impact on (say) Blacks or on women. If disparate impact is best understood as an effort to ferret out disparate treatment, that might not be a problem—at least so long as no human being, armed with a discriminatory motive, is behind its use. But if the disparate impact test is an effort to prevent something like a caste system, an algorithm that creates such an impact deserves careful scrutiny.

Different problems are presented if an algorithm uses a factor that is not race or sex as such, but that is in some sense an outgrowth of discrimination.[122] For example, a poor credit rating or a troubling arrest record might be an artifact of discrimination by human beings that predates the effort to ask the algorithm to do its predictive work. Alternatively, an algorithm might predict an outcome that may itself

[119] *Id.*

[120] *See* ELIZABETH HINTON, LESHAE HENDERSON & CINDY REED, AN UNJUST BURDEN: THE DISPARATE TREATMENT OF BLACK AMERICANS IN THE CRIMINAL JUSTICE SYSTEM 2 (2018) (summarizing decades of racial discrimination within the U.S. criminal justice system).

[121] *See* Ludwig & Mullainathan, *supra* note 44, at 82–88.

[122] *See id.* at 82.

be infected by discrimination, such as arrests for low-level offenses[123] or the promotion decisions of managers. In such cases, we even might turn out to have disparate treatment. If an algorithm predicts customer choices, and if those choices are discriminatory on the basis of, say, race, sex, disability, or age, we might again have disparate treatment. There is a risk here that algorithms might perpetuate discrimination, and even extend its reach, by using factors that are genuinely predictive, but that are products of unequal treatment.[124] They might turn discrimination into a kind of self-fulfilling prophecy. And these examples do not, of course, exhaust the potential effects of algorithms in perpetuating or creating discrimination.

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.[125] There is no general requirement of "fairness." Nonetheless, some people are keenly interested in reducing racial and other disparities—for example, in education, health care, and the criminal justice system. One of the signal virtues of algorithms is that they present the relevant tradeoffs in an unprecedentedly clear light. We might learn that if we pursue racial balance, we will simultaneously promote or sacrifice other goals, and we might be able to see, with real precision, the magnitude of the gains and the losses. We might even be able to combat discrimination through use of algorithms that identify when disparate treatment or disparate impact is occurring.[126] One advantage of the bail study is that it offers a clear illustration. Some of the tradeoffs might well be painful, but in general, it is best to know what they are.

## CONCLUSION

Inside and outside of government, the use of algorithms is often motivated by an appreciation of the limitations of human judgment. In the private and public sectors, people are often asked to make predictions under conditions of uncertainty, and their intuitions can

[123] *Id.* at 89.

[124] *See* Kleinberg et al., *Discrimination in the Age of Algorithms*, *supra* note 50, at 115–17.

[125] *See, e.g.*, Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 726 (2007).

[126] *See* Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 532–33 (2018).

lead them astray.[127] It takes a great deal of work to provide corrections.[128] It is often believed that experts can develop reliable intuitions or rely instead on statistical thinking. That is frequently true, certainly when they receive prompt feedback.[129] But as Current Offense Bias makes clear, experienced judges—here, in the literal sense—can do significantly worse than algorithms. In the administrative state, there is an insistent need for strategies that reduce noise and bias. Cost-benefit analysis can help in that enterprise, and the same is true for clear guidelines.

Algorithms are noise free, and that is an important point in their favor. Accuracy is improved when noise is reduced. To the extent that algorithms rely on statistical predictors, they will not fall prey to c-biases, such as availability bias and optimistic bias. That is also an important point in their favor. Agencies seeking to reduce noise and bias should therefore give careful consideration to the use of algorithms to the extent feasible, just as they should give careful consideration to the use of rules.

The problem of discrimination is different and complex, and I have just scratched the surface here, with reference to only 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, (4) efforts to ensure that what is predicted is not a product of discrimination, and (5) efforts to ensure racial or gender balance. For the future, (3), (4), and (5) will present many of the most important issues for the use of algorithms. For (5), and contrary to a widespread view, a primary advantage of algorithms is potential transparency: they will force people to make judgments about synergies and tradeoffs among compelling policy goals.

My central claims here, however, are narrower and (I hope) less contentious. First, algorithms eliminate noise, and that is important; to the extent that they do so, they prevent unequal treatment and reduce errors. Second, noise-free but error-prone algorithms or noise-free but

[127] For a different perspective, see generally RALPH HERTWIG, TIMOTHY J. PLESKAC & THORSTEN PACHUR, TAMING UNCERTAINTY (2019) (arguing that "uncertainty and lack of knowledge" meddle with constructions of the future, as well as with conceptions of the past).

[128] For an engaging and relevant treatment, see generally RUTH BEYTH-MAROM, SHLOMITH DEKEL, RUTH GOMBO & MOSHE SHAKED, AN ELEMENTARY APPROACH TO THINKING UNDER UNCERTAINTY (Sarah Lichtenstein, Benny Marom & Ruth Beyth-Marom trans., 1985) (discussing the psychology of decision-making under uncertainty).

[129] *See* THALER & SUNSTEIN, *supra* note 15, at 97–98.

biased algorithms are nothing to celebrate. Third, algorithms rely on statistical predictors, which means that they can counteract or even eliminate c-biases. For the administrative state, they create new and exceptionally promising opportunities.