School of Law

University of California, Davis

400 Mrak Hall Drive

Davis, CA 95616

530.752.0243

http://www.law.ucdavis.edu

UC Davis Legal Studies Research Paper Series

Research Paper No. 498

June 2016

The Racist Algorithm?

Anupam Chander

School of Law

University of California, Davis

400 Mrak Hall Drive

Davis, CA 95616

530.752.0243

http://www.law.ucdavis.edu

UC Davis Legal Studies Research Paper Series

Research Paper No. 498

June 2016

The Racist Algorithm?

Anupam Chander

Michigan Law Review

Volume 115 | Issue 6

2017

The Racist Algorithm?

Anupam Chander
University of California, Davis School of Law

THE RACIST ALGORITHM?

*Anupam Chander**

THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION. By *Frank Pasquale*. Cambridge and London: Harvard University Press. 2015. P. 218, $35.

INTRODUCTION

A pie chart satirizing Google's research and development expenditures imagines a largely tripartite division: omniscience, omnipresence, and omnipotence.[1] Uber offers its staff what it calls "God View," a real-time view of where all its users are going in a city.[2] In his new book, *The Black Box Society: The Hidden Algorithms That Control Money and Information*, Frank Pasquale[3] worries about the efforts of Silicon Valley companies to create a god in the machine. In Pasquale's forceful telling, the pie chart is not satire, but rather audacious ambition; he quotes Google's cofounder Sergey Brin, "[T]he perfect search engine would be like the mind of God" (p. 187).

We are increasingly living in a Matrix that most of us do not perceive. Pasquale is our Neo, compelling us to see the invisible digital overlords surrounding us. Pasquale does not allege the near coming of some dystopian fantasy of a networked Borg entity keen to enslave the galaxy or a Skynet

* Director, California International Law Center, Professor of Law and Martin Luther King, Jr. Hall Research Scholar, University of California, Davis School of Law. I learned much from presentations at the Intellectual Property Scholars Conference at Stanford Law School, at a faculty workshop at UC Davis School of Law, and at seminars at Berkeley Law and Tel Aviv University, and am grateful to Jacob Assaf, Ian Ayres, Michael Birnhack, Ben Blink, Steffi Bryson, Deven Desai, Chris Elmendorf, Brett Frischmann, Eric Goldman, Angela Harris, Tim Hwang, Giuseppe Labianca, Frank Pasquale, Russell Robinson, Brian Soucek, Madhavi Sunder, Tim Wilkins, and Felix Wu for very helpful suggestions and to Jennifer Reed and Nida Siddiqui for excellent research assistance, though the views expressed herein are mine alone. I am grateful as well to a Google Research Award that supported related research.

1. Chartgeist, *Google R&D Funding Breakdown*, WIRED, Nov. 2014, at 72 (note that a small sliver of research funds is left over for doodles).

2. "When new employees see the God View, they end up watching it for hours—not because they have to, but because they're just amazed by it," Uber's CEO Travis Kalanick said. Brian X. Chen, *App-Powered Car Service Leaves Cabs in the Dust, App Stars*, WIRED (Apr. 5, 2011, 7:00 AM), https://www.wired.com/2011/04/app-stars-uber/ [https://perma.cc/9UVM-RNXG]. Recently, Uber had to make restrictive changes to "God View" to settle an investigation into its privacy and security practices. *See* Kim Bellware, *Uber Settles Investigation into Creepy 'God View' Tracking Program*, HUFFINGTON POST (Jan. 6, 2016, 8:15 PM), http://www.huffingtonpost.com/entry/uber-settlement-god-view_us_568da2a6e4b0c8beacf5a46a [https://perma.cc/5XJQ-4YGL].

3. Professor of Law, University of Maryland; Affiliate Fellow, Information Society Project, Yale Law School; and Member, Council for Big Data, Ethics, and Society.

bent on terminating humanity. Rather, he worries that decisionmaking on everything, from credit to employment to investments to even dating, is passing from humans to computers.[4] And these computers are remote and invisible, their algorithms protected from scrutiny by trade secret law, invisibly and relentlessly manipulating us for the benefit of corporate profit or worse (pp. 6–14). Pasquale shows that corporations often rebuff efforts to examine the algorithms they employ, and the law abets corporations in this task (Chapter Five).

Pasquale is part of a line of recent scholarship attacking the increasing role of automated algorithms in our lives—indeed, legal scholars are increasingly sounding the alarm on this unfettered algorithmic control. Jonathan Zittrain worries that a company like Facebook could even decide an election without anyone ever finding out.[5] Ryan Calo warns that companies may be manipulating us through advertising.[6] Call this the problem of *algorithmic manipulation*.[7]

I will argue that despite his careful and important account, Pasquale's "black box society" frame lends itself to a misdiagnosis of the discrimination problem likely to lie in algorithmic decisionmaking. This misdiagnosis leads to the wrong prescription—namely, an often-quixotic search for algorithmic transparency.[8] Furthermore, the transparency that Pasquale's argument can be read to support is the wrong sort: a *transparency in the design* of the algorithm. (I should make clear that Pasquale himself is more nuanced, calling for a discussion of the kinds of transparency we should demand; he asks: "How much does the black box firm have to reveal? To whom must it reveal it? And how fast . . . ?" (p. 142).) Even a transparent, facially neutral algorithm can still produce discriminatory results.[9] What we need instead is a *transparency of inputs and results*, which allows us to see that the algorithm is

4. Danielle Citron has made a similar observation about the increasing role of computers in decisionmaking by public sector entities. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008).

5. Jonathan Zittrain, *Facebook Could Decide an Election Without Anyone Ever Finding Out*, NEW REPUBLIC (June 1, 2014), https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering (on file with *Michigan Law Review*).

6. Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2014).

7. Intentional discrimination is not the only complaint cited against algorithmic decisionmaking. Anne Cheung observes that Google's autocomplete function can sometimes result in distressing results. She argues that Google and others should be liable for defamatory autocompletes that are algorithmically generated if they fail to take down the particular autocomplete even after being notified of its allegedly defamatory nature. Anne S.Y. Cheung, *Defaming by Suggestion: Searching for Search Engine Liability in the Autocomplete Era*, *in* COMPARATIVE PERSPECTIVES ON THE FUNDAMENTAL FREEDOM OF EXPRESSION 467 (András Koltay ed., 2015).

8. Pasquale recognizes the possibility of algorithms too complex to understand—what he calls the "sweet mystery of machine learning." Frank Pasquale, *Bittersweet Mysteries of Machine Learning (A Provocation)*, LONDON SCH. ECON. & POL. SCI.: MEDIA POL'Y PROJECT BLOG (Feb. 5, 2016), http://blogs.lse.ac.uk/mediapolicyproject/2016/02/05/bittersweet-mysteries-of-machine-learning-a-provocation/ [https://perma.cc/XSS9-2D58].

9. *See* discussion *infra* Part II.

generating discriminatory impact. If we know that the results of an algorithm are systematically discriminatory, then we know enough to seek to redesign the algorithm or to distrust its results. The distinction is similar to the evidentiary difference between demonstrating disparate treatment and demonstrating disparate impact.[10] My central claim is this: if we believe that the real-world facts, on which algorithms are trained and operate, are deeply suffused with invidious discrimination, then our prescription to the problem of racist or sexist[11] algorithms is *algorithmic affirmative action*. Thus, the problem is not the black box, which is often more neutral than the human decisionmaker it replaces, but the real world on which it operates. We must design our algorithms for a world permeated with the legacy of discriminations past and the reality of discriminations present.

The importance of getting this right is clear. Facebook now owns a patent on a process by which a user can be denied a loan because of the creditworthiness of his or her friends.[12] IBM purports to offer an algorithm that can distinguish refugee from terrorist, "the sheep from the wolves."[13]

---

10. While casting some doubt on the continuing validity of disparate impact justifications for civil rights remedies in its 2009 decision in *Ricci v. DeStefano*, 557 U.S. 557 (2009), the Supreme Court in its latest ruling on issues of race upheld a disparate-impact-based theory of civil rights violations under the Fair Housing Act. There the Court distinguished disparate treatment and disparate impact theories as follows: "In contrast to a disparate-treatment case, where a 'plaintiff must establish that the defendant had a discriminatory intent or motive,' a plaintiff bringing a disparate-impact claim challenges practices that have a 'disproportionately adverse effect on minorities' and are otherwise unjustified by a legitimate rationale." Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2533 (2015) (quoting *Ricci*, 557 U.S. at 577). The Court warned that remedying disparate impacts through explicit "racial quotas" would raise "serious constitutional concerns." *Id.* at 2523. On the contemporary use of disparate impact in classifications, see Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After* Inclusive Communities, 101 Cornell L. Rev. 1115 (2016).

11. For economy only, I will speak in terms of race- and gender-based discrimination, though I intend my argument to be more broadly applicable across the array of prohibited discrimination criteria. For an example of work examining age-based discrimination (in the context of credit scoring), see Faisal Kamiran & Toon Calders, Classifying Without Discriminating (2009), http://ieeexplore.ieee.org/ielx5/4850420/4909154/04909197.pdf?tp=&arnumber=4909197&isnumber=4909154 [https://perma.cc/WZ3D-ZLVA].

12. Hazel Sheffield, *Facebook's New Patent Could Mean You Are Denied a Loan Because of Your Friends*, Independent (Sept. 2, 2015), http://www.independent.co.uk/news/business/news/facebooks-new-patent-could-mean-you-are-denied-a-loan-because-of-your-friends-10482622.html [https://perma.cc/2R8K-EE5Q]; Mark Sullivan, *Facebook Patents Technology to Help Lenders Discriminate Against Borrowers Based on Social Connections*, VentureBeat (Aug. 4, 2015, 12:15 PM), http://venturebeat.com/2015/08/04/facebook-patents-technology-to-help-lenders-discriminate-against-borrowers-based-on-social-connections/ [https://perma.cc/C2LK-VPLY].

13. Patrick Tucker, *Refugee or Terrorist? IBM Thinks Its Software Has the Answer*, Def. One (Jan. 27, 2016), http://www.defenseone.com/technology/2016/01/refugee-or-terrorist-ibm-thinks-its-software-has-answer/125484/ [https://perma.cc/S27J-KR6S]. We also see the emergence of preemptive policing algorithms, which may have the effect of increasingly targeting minority neighborhoods, thus subjecting minorities to greater surveillance, and perhaps greater risk of accidental or wrongful use of police force. *See* Kelly K. Koss, Note, *Leveraging Predictive Policing Algorithms to Restore Fourth Amendment Protections in High-Crime Areas in*

Retailers can increasingly target certain shoppers for discounts.[14] Law enforcement officers are using "predictive policing" algorithms to identify "hot people" who might have a greater propensity to commit crime.[15] Judges are employing algorithms in sentencing.[16] As Pasquale describes, Google's search and Facebook's presentation algorithms determine what information we see (p. 82). The possibilities of discriminatory manipulation are legion.[17] Pasquale worries that the rise of algorithmic decisionmaking will make racism and other discrimination even more difficult to ferret out, hidden behind subtle manipulations that are nearly impossible to discern for ordinary citizens not privy to the internal computer code (p. 38). "*It [c]reates [i]nvisible [p]owers*," he warns (p. 193).

Pasquale's warning comes at a time when the #BlackLivesMatter campaign and other recent events have made the reality of racial- and gender-based discrimination in our society painfully clear.[18] In one famous experiment, job applicants with white-sounding names, such as Emily, received 50 percent more callbacks than those with African American–sounding names, such as Lakisha.[19] A study of emails sent to mortgage loan originators asking for loans found that African American–sounding names effectively reduced an applicant's credit score by 71 points (on a scale going up to 750).[20] A 2012 federal government report found that both African Americans and Asian Americans were shown 17.7 percent fewer homes than equally

---

a Post-Wardlow *World*, 90 CHI.-KENT L. REV. 301, 321 (2015) (describing how high-crime areas are disproportionately low-income, minority neighborhoods across the United States).

14. Joseph Turow & Lee McGuigan, *Retailing and Social Discrimination: The New Normal?*, *in* DATA AND DISCRIMINATION: COLLECTED ESSAYS 27, 29 (Seeta Peña Gangadharan ed., 2014), https://www.ftc.gov/system/files/documents/public_comments/2014/10/00078-92938.pdf [https://perma.cc/K7B7-P6X8].

15. John Eligon & Timothy Williams, *Police Program Aims to Pinpoint Those Most Likely to Commit Crimes*, N.Y. TIMES (Sept. 24, 2015), http://www.nytimes.com/2015/09/25/us/police-program-aims-to-pinpoint-those-most-likely-to-commit-crimes.html (on file with *Michigan Law Review*).

16. *E.g.*, State v. Loomis, 881 N.W.2d 749, 754–55 (Wis. 2016) (evaluating the use of an algorithm that purports to predict whether an criminal defendant will reoffend).

17. The possibility of manipulation is heightened because many do not even recognize that their experiences with a particular process are determined by a human-coded algorithm.

18. *See e.g.*, Mario L. Barnes et al., *A Post-Race Equal Protection?*, 98 GEO. L.J. 967, 982–92 (2010) (collecting statistics demonstrating persistence of racism).

19. Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991, 997–99 (2004).

20. Andrew Hanson et al., *Discrimination in Mortgage Lending: Evidence from a Correspondence Experiment*, 92 J. URB. ECON. 48, 62 (2016) ("We find that [mortgage loan originators] . . . are less likely to respond to inquiries from clients with African American names than they are to clients with white names. We also find that [originators] . . . are likely to write a preferential e-mail to white clients."); *cf.* John Leland, *Baltimore Finds Subprime Crisis Snags Women*, N.Y. TIMES, Jan. 15, 2008, at A1, http:// www.nytimes.com/2008/01/15/us/15mortgage.html (disproportionate amount of Baltimore's subprime home loans have gone to women regardless of their income or credit scores).

qualified white Americans.[21] A 2015 federal government study found yet other invidious discrimination in housing: housing providers tell deaf or hard-of-hearing homeseekers about fewer units than similar homeseekers who are not deaf.[22] Homeseekers who use wheelchairs are more likely to be denied an appointment to view rental housing in buildings with accessible units, and when given an appointment, are less likely to be shown suitable housing units than homeseekers who are ambulatory.[23] In a society where discrimination affects opportunities in innumerable ways, we must worry about the migration of discrimination to decisionmaking by algorithm.

This Review proceeds as follows. Part I reviews Pasquale's argument that our emerging black box society will increase discriminatory manipulations. It argues that, contrary to Pasquale's argument, instead of seeing algorithms as likely to *increase* intentional discrimination, the law has turned to algorithms to *reduce* the invidious discriminations that result from human decisionmakers with unfettered discretion. Through the example of sentencing guidelines, this Part demonstrates that law has preferred highly specified algorithmic decisionmaking in order to reduce the discriminatory results of open-ended human judgment. Part II argues that because of the real-world discrimination upon which the algorithms learn and operate, discrimination is still likely to emerge from automated algorithms that are designed in racially or gender-neutral fashion. Part III introduces the remedy of algorithmic affirmative action to combat the problem of viral discrimination—designing algorithms in race- and gender-conscious ways to account for existing discrimination lurking in the data.

I. Algorithmic Manipulation

Pasquale deploys two striking Platonic metaphors to illustrate his concerns. First, he sees the data industry as wearing a ring of invisibility: "Black box insiders are protected as if they are wearing a Ring of Gyges—which grants its wearers invisibility but, Plato warns us in *The Republic*, is also an open invitation to bad behavior" (p. 190). Second, Pasquale posits the rest of us ordinary people as prisoners in Plato's allegory of the cave, forced to stare at a stony wall "flickering shadows cast by a fire behind them" (p. 190). Pasqaule concludes:

> [We prisoners in the cave] cannot comprehend the actions, let alone the agenda, of those who create the images that are all [we] know of reality. Like those who are content to use black box technology without

21. Margaret Austin Turner et al., U.S. Dep't of Hous. & Urban Dev., Housing Discrimination Against Racial and Ethnic Minorities 2012, at xvii (2013), http://www.huduser.gov/portal/Publications/pdf/HUD-514_HDS2012.pdf [https://perma.cc/NJU4-L4B6].

22. Diane K. Levy et al., U.S. Dep't of Hous. & Urban Dev., Discrimination in the Rental Housing Market Against People Who Are Deaf and People Who Use Wheelchairs: National Study Findings 37–42 (2015), https://www.huduser.gov/portal/sites/default/files/pdf/housing_discrimination_disability.pdf [https://perma.cc/4N6D-GYVK].

23. *Id.* at 42–52.

> understanding it, [we] can see mesmerizing results, but [we] have no way to protect [ourselves] from manipulation or exploitation (p. 190).

Given the persistence of widespread racial and gender discrimination in the twenty-first century, should we not expect algorithms often programmed by racist and sexist programmers to manipulate us towards accepting racist and sexist decisions? Are programmers likely to manipulate algorithms to exacerbate existing discrimination in society? For a half-dozen reasons, I believe the answer is no. (Pasquale, I should note, does not suggest either rogue programmers or malign bosses, but the concern about algorithmic manipulation might be interpreted that way.)

First, because much of societal discrimination is subconscious or unconscious, it is less likely to be encoded into automated algorithms than the human decisionmakers that the algorithms replace.[24] Much of recent research into racial bias has moved toward exposing its existence without focusing on whether it is conscious or not. Implicit association testing has revealed the prevalence of bias across the community.[25] As Jerry Kang writes, "[W]e may all be infected in ways we cannot admit, even to ourselves."[26] Research focused on implicit bias often posits that the bias is unconscious.[27] The Supreme Court in 2015 recognized that "unconscious prejudices" can motivate discrimination.[28] Unconscious discrimination is far less likely to manifest itself through the process of programming than through the process of decisionmaking. Programming requires a step-by-step writing

---

24. Susan T. Fiske & Ann Marie Russell, *Cognitive Processes*, *in* THE SAGE HANDBOOK OF PREJUDICE, STEREOTYPING AND DISCRIMINATION 115, 124 (John F. Dovidio et al. eds., 2010) ("Subtle, unexamined stereotyping is more automatic, ambiguous, and ambivalent than common sense would assume.").

25. *See* Audrey J. Lee, *Unconscious Bias Theory in Employment Discrimination Litigation*, 40 HARV. C.R.-C.L. L. REV. 481, 482 (2005) ("A burgeoning body of social science literature has empirically demonstrated the existence and prevalence of unconscious bias in today's society."). *See generally* Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CALIF. L. REV. 945 (2006); Justin D. Levinson, *Forgotten Racial Equality: Implicit Bias, Decisionmaking, and Misremembering*, 57 DUKE L.J. 345, 352 (2007) ("[I]mplicit biases are real, pervasive, and difficult to change.").

26. Jerry Kang, *Trojan Horses of Race*, 118 HARV. L. REV. 1489, 1496 (2005). Kang cites a landmark article by Charles Lawrence, who wrote that "the illness of racism infects almost everyone." Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317, 321 (1987).

27. Consider the description offered recently by Song Richardson and Phillip Goff: "We use the term implicit racial biases to refer both to unconscious stereotypes (beliefs about social groups) and attitudes (feelings, either positive or negative, about social groups). Implicit stereotypes and attitudes result from the practice we get associating groups (e.g., blacks) with traits (e.g., criminality)." L. Song Richardson & Phillip Atiba Goff, *Implicit Racial Bias in Public Defender Triage*, 122 YALE L.J. 2626, 2630 (2013). Richardson and Goff suggest that implicit bias stems from prevalent media and other cultural portrayals: "This practice stems from repeated exposures to cultural stereotypes that are ubiquitous within a given society." *Id.*

28. Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2511–12 (2015) ("Recognition of disparate-impact liability under the FHA plays an important role in uncovering discriminatory intent: it permits plaintiffs to counteract unconscious prejudices and disguised animus that escape easy classification as disparate treatment.").

process that depends on a conscious understanding of what is sought.[29] Not only must the programmer instruct the computer with great precision, but modern programming practices also require the programmer to document (or annotate) what the program is doing.[30] Because of a programming process that requires both writing down explicit instructions and documenting what particular code does, unconscious or subconscious discrimination is less likely to manifest itself in computer programming than in human decisionmaking.

Return to Bertrand and Mullainathan's famous resume experiment.[31] Would a computer share the same biases in deciding whom to call back, preferring Emily and Greg over equally qualified Lakisha and Jamal? That depends, of course, on whether the computer could divine the race of the applicants from their name. It seems hard to imagine that programmers would teach an algorithm to distinguish race from the name of an applicant. Given laws barring racial discrimination in employment, such programming would smack of intentional racial discrimination and would prove devastating in a trial for any discrimination claim.[32] A computer program would thus not likely guess the race of the applicant from the names Emily, Greg, Lakisha, and Jamal—a guess that many humans would make, perhaps subconsciously. Callbacks by computer algorithm seem less likely to discriminate than human resources personnel, with their hidden and often unconscious biases.

Second, even for programmers or companies who intend to discriminate, the process of coding itself is likely to cause programmers to shy away from actually encoding the discrimination. Even absent compelled disclosure through litigation, there is the danger that a hard-coded discrimination will be revealed later by hackers or by insiders disgusted by the discrimination. Moreover, because code writing is likely to involve teams of programmers sharing code, with different persons reviewing and debugging code, consciously coding discrimination will likely require obtaining the cooperation of multiple persons, which is likely to be a fraught task. None of this denies the existence of racist and sexist programmers. Recently, Mark Zuckerberg repudiated some Facebook employees who replaced declarations of "Black Lives Matter" on company walls with "All Lives Matter."[33]

---

29. *Computer Programming*, U.R.I. Dep't Computer Sci. & Stat., http://homepage.cs.uri.edu/book/programming/programming.htm [https://perma.cc/K8Y2-GND9].

30. *Id.* (breaking down programming into five steps: defining the problem, planning the solution, coding the program, testing the program, and documenting the program).

31. Bertrand & Mullainathan, *supra* note 19 and accompanying text.

32. 42 U.S.C. § 2000e–2(a) (2012) ("It shall be an unlawful employment practice for an employer . . . to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin . . . .").

33. Michael Nunez, *Mark Zuckerberg Asks Racist Facebook Employees to Stop Crossing Out Black Lives Matter Slogans*, Gizmodo.com (Feb. 25, 2016, 12:42 PM), http://gizmodo.com/

Third, even if corporations engage in other kinds of wrongdoing, that does not mean that they are likely to intentionally manipulate algorithms to invidiously discriminate. Christian Sandvig and his coauthors argue that history indicates that we should be distrustful of some of the major operators of algorithms:

> Almost every major operator of an Internet platform, including Google, Twitter, Facebook, Microsoft, and Apple, has already been investigated by the U.S. government for violations that include anti-competitive behavior, deceptive business practices, failing to protect the personal information of consumers, failing to honor promises made to consumers about their own data, and charging customers for purchases that they did not authorize.[34]

The history that Sandvig and his coauthors cite, however, does not reveal intentional racial or gender discrimination. Certainly, many have rightly pointed to concerns about the demographics of high-tech workplaces, where certain minorities and women are not represented as well as they are in the general population.[35]

Fourth, the process must be compared with non-algorithmic decisionmaking—that is, decisionmaking by human beings. The ultimate black box is the human mind. Even where decisions are made by committee, the deliberations are typically not recorded except in highly selective minutes, and the members of the committee often pledge to hold discussions secret. Prejudices acted upon in this black box never have to be written down. Consider Zittrain's concern about Facebook's manipulation of its newsfeed to favor one candidate or another in an election.[36] The decision as to what constitutes news has previously been the realm of editorial desks in television stations, magazines, and newspapers—hardly a guarantee of political neutrality. How many elections have actually been swayed by human editorial decisions as to what to reveal or highlight? Even one of the most trusted newspapers in the country, the *New York Times*, withheld "a blockbuster story . . . about a secret Bush administration program to eavesdrop on Americans without warrants" until after the November 2004 presidential election.[37]

Fifth, if human beings act on stereotypes formed through a process of statistical discrimination, automated algorithms acting on a richer information environment may not be subject to similar individually erroneous statistical discrimination. Lior Strahilevitz has argued that much of the

---

mark-zuckerberg-asks-racist-facebook-employees-to-stop-1761272768 [https://perma.cc/JT2U-FQTS].

34. Christian Sandvig et al., *An Algorithm Audit*, *in* Data and Discrimination: Collected Essays, *supra* note 14, at 6, 7–8.

35. Patricia Leigh Brown, *Silicon Valley, Seeking Diversity, Focuses on Blacks*, N.Y. Times (Sept. 3, 2015), http://www.nytimes.com/2015/09/04/technology/silicon-valley-seeking-diversity-focuses-on-blacks.html (on file with *Michigan Law Review*).

36. Zittrain, *supra* note 5.

37. *See* Margaret Sullivan, *Lessons in a Surveillance Drama Redux*, N.Y. Times, Nov. 9, 2013, at 12.

discrimination that prevails in contemporary society is a result of statistical discrimination, where, for example, "in the absence of accurate information about individuals' criminal histories, employers who are interested in weeding out those with criminal records will rely instead on racial and gender proxies."[38] A more comprehensive portrait of an individual might allow algorithms to avoid falling prey to racial or gender stereotypes (though, as I discuss in Part II below, this is hardly inevitable).

Finally, worries about the black box of human decisionmaking have led to algorithmic turns in the past. When Facebook was accused of favoring liberal political views in the news stories it identified as "trending," it sought to clearly demonstrate its commitment to fairness by firing its news editors and replacing them fully with an automated algorithm.[39]

Consider the history of the federal sentencing guidelines, to which we now turn. In 1976, Senator Edward Kennedy declared the sentencing disparities resulting from the indeterminate sentencing of federal crimes by judges "a national scandal."[40] In the volume of *Judicature* published the year before, state judge Joseph C. Howard observed that "[f]indings in other fields converge on the fact that cultural distance tends to determine attitude and tolerance, and the greater the distance in this respect, the greater the tendency for imagination and bias to influence judgment and decisions."[41] Criminal justice scholar Albert Alschuler echoed the worry: "Whenever discretion is granted, it will be abused. In some instances, individual differences in culpability will be less important than differences in race, class, lifestyle and other irrelevancies."[42] As Kate Stith and Steven Koh describe in their history of this period, "[D]iscrepancy in sentences was said to be fundamentally at odds with ideals of equality and the rule of law. In particular, permitting judges and parole officials to exercise unguided discretion resulted in 'unwarranted

38. *See* Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. CHI. L. REV. 363, 364 (2008).

39. Annalee Newitz, *Facebook Fires Human Editors, Algorithm Immediately Posts Fake News*, ARS TECHNICA (Aug. 29, 2016, 2:20 PM), http://arstechnica.com/business/2016/08/facebook-fires-human-editors-algorithm-immediately-posts-fake-news/ [https://perma.cc/8BT6-LW9C]; *Search FYI: An Update to Trending*, FACEBOOK NEWSROOM (Aug. 26, 2016) http://newsroom.fb.com/news/2016/08/search-fyi-an-update-to-trending/ [https://perma.cc/Y693-RJAF] (Facebook explains changes where human editors will "make fewer individual decisions about topics"); *see also* Mike Isaac, *Facebook, Facing Bias Claims, Shows How Editors and Algorithms Guide News*, N.Y. TIMES (May 12, 2016), http://www.nytimes.com/2016/05/13/technology/facebook-guidelines-trending-topics.html (on file with *Michigan Law Review*) (providing background on the criticism Facebook was receiving); FACEBOOK, TRENDING REVIEW GUIDELINES (2016), https://fbnewsroomus.files.wordpress.com/2016/05/full-trending-review-guidelines.pdf [https://perma.cc/MP29-GB5U] (Facebook's public news team guidelines, released after the public criticism of bias).

40. Edward M. Kennedy, *Criminal Sentencing: A Game of Chance*, 60 JUDICATURE 208, 210 (1976).

41. Joseph C. Howard, *Racial Discrimination in Sentencing*, 59 JUDICATURE 121, 122 (1975).

42. Albert W. Alschuler, *Sentencing Reform and Prosecutorial Power: A Critique of Recent Proposals for "Fixed" and "Presumptive" Sentencing*, 126 U. PA. L. REV. 550, 563 (1978).

disparity' (including alleged bias against minorities) in criminal sentences."[43] Motivated in part by concerns over sentencing disparities resulting from the largely unfettered sentencing discretion granted to federal judges, Congress, in 1984, created the United States Sentencing Commission, charging it to promulgate sentencing policies that would help "avoid unwarranted sentencing disparities among defendants with similar records who have been found guilty of similar criminal conduct."[44] The Sentencing Guidelines made the process of determining a sentence more algorithmic, based on the nature of the offense and various other features of the crime and the criminal. Presaging the concern about viral discrimination hidden in data, Congress had heard concerns that the sentencing algorithm itself might unintentionally utilize race through other factors that might serve as hidden proxies for race.[45] Today, the Federal Sentencing Guidelines Manual spans some 600 pages, complete with extensive tables and indices.[46]

Despite these efforts, sentencing based on ranges determined by algorithm did not end discrimination and might have made the lives of African American convicted criminals (like white American convicted criminals) worse through lengthier mandatory minimums. Experience with the Sentencing Guidelines does not suggest an unequivocal story of human invidious discriminations restrained by highly specified algorithms. Stith and Koh worry that by reducing human discretion, the Sentencing Guidelines made minorities worse off because their individual circumstances could not be properly considered.[47] Indeed, the Guidelines themselves had rules that treated crack cocaine—more popular with African American drug addicts—more harshly than powder cocaine—more popular with white American

43. Kate Stith & Steve Y. Koh, *The Politics of Sentencing Reform: The Legislative History of the Federal Sentencing Guidelines*, 28 WAKE FOREST L. REV. 223, 227 (1993).

44. Sentencing Reform Act of 1984, Pub. L. No. 98-473, tit. II, ch. 2, 98 Stat. 1987 (codified as amended in scattered sections of 18 and 28 U.S.C.).

45. Before a Congressional Subcommittee on Criminal Laws and Procedures, attorney Pierce O'Donnell testified along with law professors Michael Churgin and Dennis Curtis:

> Under § 994(d), the Sentencing Commission is required to consider certain factors in classifying categories of offenders for purposes of its sentencing guidelines. Some of these factors include the offender's education, family and community ties, vocational skills and previous employment records. The Subcommittee, we are sure, is aware of the care which must be used in employing such considerations. Not only are there serious doubts about the utility of some of these factors in making assessments of risk of recidivism. There is also the potential for inadvertent discrimination on the basis of race and income.

Reform of the Federal Criminal Laws: Hearing on S. 1473 Before the Subcomm. On Criminal Laws & Procedures of the S. Comm. on the Judiciary, 95th Cong. 8913 (1977). Harvard Law Professor Alan Dershowitz agreed, declaring that he "would hope that [the members appointed to the Sentencing Commission] would reflect the wide and rich diversity ethnically and in terms of gender and race that we have in this country." *Id.* at 9051.

46. *See* U.S. SENTENCING GUIDELINES MANUAL (U.S. SENTENCING COMM'N 2015).

47. Stith & Koh, *supra* note 43, at 287.

drug addicts.[48] Nearly two decades after their adoption, Federal District Judge Myron Thompson complained, "If the 600-plus pages of the most recent set of sentencing guidelines have taught us anything, it is that punishment cannot be reduced to an algorithm."[49]

In a recent case, an individual facing a six-year prison term challenged the sentencing judge's use of an algorithmic tool called "COMPAS," arguing that the algorithm was based on group data, that the algorithm took gender into account, and that the algorithm's proprietary nature prevented him from challenging its validity.[50] The Wisconsin Supreme Court upheld the use of the algorithm, within narrow parameters, arguing that "consideration of a COMPAS risk assessment at sentencing along with other supporting factors is helpful in providing the sentencing court with as much information as possible in order to arrive at an individualized sentence."[51] With respect to gender, the court concluded that "COMPAS's use of gender promotes accuracy that ultimately inures to the benefit of the justice system including defendants."[52] Meanwhile, a ProPublica investigation challenged COMPAS as "likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants."[53]

In the context of decisions made by governments that are subject to constitutional obligations of due process, there is reason to allow individuals to review any algorithms that are crucial parts of that decisionmaking.[54] In such circumstances, the government's interest in protecting trade secrets seems an inadequate reason to refuse to inform citizens of why the government took decisions that seriously impact their lives. Perhaps concerns

48. David A. Sklansky, Essay, *Cocaine, Race, and Equal Protection*, 47 STAN. L. REV. 1283, 1287–89 (1995) ("The 100:1 ratio between the sentencing thresholds for powder cocaine and crack is mirrored in the Sentencing Guidelines promulgated in 1987.").

49. Myron H. Thompson, *Sentencing and Sensibility*, N.Y. TIMES, (Jan. 21, 2005), http://www.nytimes.com/2005/01/21/opinion/sentencing-and-sensibility.html [https://perma.cc/BSZ6-YAP6].

50. State v. Loomis, 881 N.W.2d 749, 754–57 (Wis. 2016).

51. *Id.* at 765.

52. *Id.* at 767.

53. *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/5HBK-MUJR]. The company behind COMPAS has contested this claim. *Northpointe's Response to ProPublica: Demonstrating Accuracy Equity and Predictive Party*, NORTHPOINTE, http://www.northpointeinc.com/northpointe-analysis [https://perma.cc/L2PP-VJ5D]. ProPublica's reporters stand by their original claim. Julia Angwin & Jeff Larson, *ProPublica Responds to Company's Critique of Machine Bias Story*, PROPUBLICA (July 29, 2016, 11:56 AM), https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story [https://perma.cc/6S29-98F4].

54. The European Union's General Data Protection Regulation, set to take effect in 2018, offers a broader promise of transparency: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." Commission Regulation 2016/679, art 22, para. 1, 2016 O.J. (L 119) 1, 46 (EU).

about manipulation if the algorithm is disclosed might justify keeping a particular algorithm secret—for example, an algorithm that singles out whom to scrutinize more carefully in an airport screening process.[55]

Pasquale recognizes that institutions have, at times, turned to algorithms to replace biased human decisionmakers. He argues that this algorithmic turn has, however, proven inadequate, offering the example of finance: "Algorithmic methods of reducing judgment to a series of steps were supposed to rationalize finance, replacing self-serving or biased intermediaries with sound decision frameworks. And they did reduce some inefficiencies. But they also ended up firmly building in some dubious old patterns of credit castes and corporate unaccountability" (p. 15). Pasquale's worry seems well-placed with respect to the adoption of sentencing guidelines—the formulation of the new algorithm itself may have solved some problems of subconscious bias among judges, but created others, with a possible overall adverse impact on minorities.

Pasquale raises another critical concern with the algorithmic turn: when algorithms replace human decisionmaking, algorithms give the decisionmaking "a patina of inevitability" (p. 15), and indeed a patina of fairness. Algorithms can make decisionmaking seem fair precisely because computers are logical entities which should not be infected by all-too-human bias. But that would be an unwarranted assumption, as we now discuss.

II. Viral Discrimination

Can a search engine peer into our souls? Consider one account: "Within the confines of a search bar you can ask questions or express opinions you would never admit to in public. Our most popular searches are, to some degree, an uncensored chronicle of what, as a society, we're thinking but not necessarily saying."[56] Google's autocomplete function uses prior searches of others to help predict what you are searching for,[57] and in this unexpected way, becomes a window into the questions that large numbers of people are asking "when they think no-one is looking."[58] The United Nations cleverly utilized Google's autocomplete function to reflect back to us our own sexist

55. *E.g.*, Ethan Zuckerman, *TSA Pre-Check, Fairness and Opaque Algorithms*, . . . My Heart's in Accra (Sept. 5, 2012), http://www.ethanzuckerman.com/blog/2012/09/05/tsa-pre-check-fairness-and-opaque-algorithms/ [https://perma.cc/P6MM-FRF9] ("[I]f we discover that flying 20 times in a year is a key factor for clearance into the program, it's easy to imagine an adversary flying 20 times with clean hand baggage and a bomb smuggled on the 21st flight.").

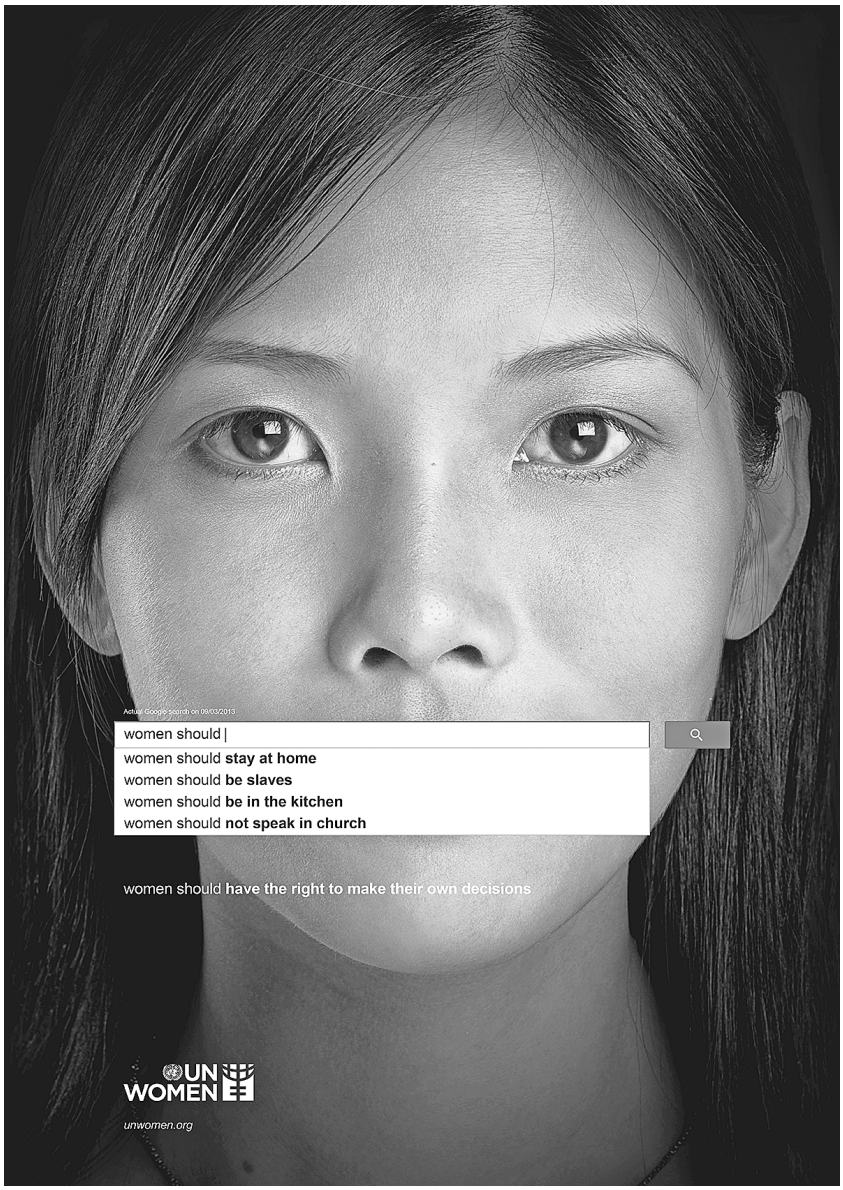
56. Arwa Mahdawi, *Google's Autocomplete Spells Out Our Darkest Thoughts*, Guardian (Oct. 22, 2013, 5:39 PM), http://www.theguardian.com/commentisfree/2013/oct/22/google-autocomplete-un-women-ad-discrimination-algorithm [https://perma.cc/H6VT-2SXP].

57. *Id.*

58. *New UN Campaigns Use Google Search Results to Reveal Prevalence of Sexism and Homophobia*, United Nations Hum. Rts. Off. High Commissioner (Oct. 31, 2013), http://www.ohchr.org/EN/NewsEvents/Pages/NewUNcampaignsuseGooglesearch.aspx#sthash.Gk1la3Y0.dpuf [https://perma.cc/H6UM-UHVK] ("UN Women recently launched a clever ad campaign designed to highlight the prevalence of sexist attitudes.").

attitudes towards women via an advertising campaign, an example of which appears below in Figure 1.

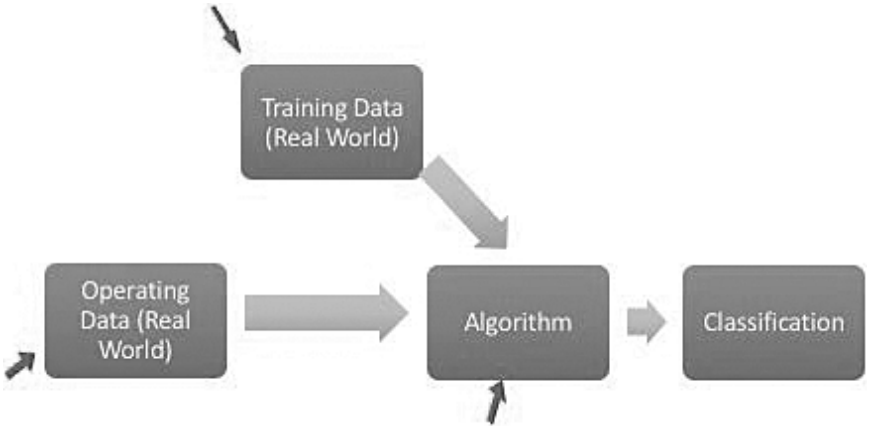
FIGURE 1. UN WOMEN AD CAMPAIGN USING GOOGLE'S AUTOCOMPLETE FUNCTION[59]



59. *UN Women Ad Series Reveals Widespread Sexism*, UN WOMEN (Oct. 21, 2013), http://www.unwomen.org/en/news/stories/2013/10/women-should-ads [https://perma.cc/L9ZU-RX2G].

The United Nations' advertisements demonstrate how a neutral algorithm can produce sexist results simply because it responds to inputs from sexist people. More generally, ostensibly neutral algorithms can produce results that reflect the prejudices of society. Thus, even if algorithms are less likely than the human decisionmakers they replace to be afflicted by prejudice, algorithms can still further entrench discrimination through other means. Even facially neutral algorithms will produce discriminatory results because they train and operate on the real world of pervasive discrimination.

FIGURE 2. ROUTES OF INFECTION IN ALGORITHMIC DECISIONMAKING



Pasquale's manipulative algorithm critique contrasts with another emerging critique—that algorithms simply compound the errors of the past (call this the problem of *viral discrimination*). Algorithms trained or operated on a real-world data set that necessarily reflects existing discrimination may well replicate that discrimination.

Figure 2 shows the routes of infection in an algorithm's decisionmaking—from training data and operating data to the algorithm itself. As Faisal Kamiran and Toon Calders write, "Classification models are trained on the historical data for the prediction of the class labels of unknown data samples. Often, however, the historical data is biased towards certain groups or classes of objects."[60] Solon Barocas and Andrew Seibst demonstrate a variety of mechanisms through which algorithmic decisionmaking in employment can lead to disparate impact against protected classes.[61] For example, algorithms might utilize ostensibly neutral data, but that data may turn out to be subject to what Cynthia Dwork and her colleagues have called "redundant encodings," where membership in a particular class is encoded in other data

60. Kamiran & Calders, *supra* note 11, at 1.

61. *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).

that can seem, on its face, neutral.[62] With redundant encodings, race or gender can even be inferred from datasets that do not explicitly collect information on race and gender. Latanya Sweeney has demonstrated that automated algorithms can generate racially problematic outcomes even if that was not the intent of the algorithms' programmers.[63] Sweeney has established that algorithmic online advertisements can unintentionally propagate discrimination, with African American–sounding names more likely to generate advertisements that relate to arrest records than names typically associated with white Americans.[64] Such discrimination arises not from the racist intentions of the advertising algorithms' programmers, but from the algorithms' natural operation in the real world.[65]

Pasquale is sensitive to the problem of bad data, rather than intentionally bad algorithms. He writes:

> After Sweeney released her findings, several explanations for her results were proposed. Perhaps someone had deliberately programmed "arrest" results to appear with names associated with blacks? That would be intentional discrimination, and Instant Checkmate and Google both vehemently denied it. On the other hand, let us suppose that (for whatever reasons) web searchers tended to click on Instant Checkmate ads more often when names associated with blacks had "arrest" associations, rather than more neutral ones. In that case, the programmer behind the ad-matching engine could say that all it is doing is optimizing for clicks—it is agnostic about people's reasons for clicking. It presents itself as a cultural voting machine, merely registering, rather than creating, perceptions (p. 39).

He notes that "without access to the underlying coding and data," it is difficult to evaluate why the ads exhibited racial bias (p. 39).

Economists have theorized how discrimination might arise even from decisionmakers not themselves motivated by invidious prejudices. A half-century ago, economists began to ask how racism or sexism can persist in a world of profit-maximizing employers. If employers are profit maximizers, "non-discriminatory employers, . . . would drive out the others," who would be paying unsustainable premiums to continue their discriminatory hiring practices.[66] Edmund Phelps and Kenneth Arrow offered one explanation:

---

62. Cynthia Dwork et al., *Fairness Through Awareness*, *in* Innovations in Theoretical Computer Science 2012, at 214 (2012), http://dl.acm.org/citation.cfm?id=2090236&preflayout=flat [https://perma.cc/S6EB-HG3U].

63. Latanya Sweeney, *Discrimination in Online Ad Delivery*, Comm. ACM, May 2013, at 44, http://cacm.acm.org/magazines/2013/5/163753-discrimination-in-online-ad-delivery/ [https://perma.cc/N6NZ-H7R3].

64. *Id.* at 50–52.

65. *Id.* at 52.

66. Kenneth Arrow describes this in economic terms: "If the members of the two races, after adjusting for observable differences in human capital and the like, received different wages or were charged different prices in commodity or credit markets, an arbitrage possibility would be created which would be wiped out by competition." Kenneth J. Arrow, *What Has Economics to Say About Racial Discrimination?*, J. Econ. Persp., Spring 1998, at 91, 94–95.

statistical discrimination.[67] Because race or gender might be statistically associated with an unobservable trait—such as worker productivity or propensity to remain in the labor market—profit-maximizing employers might discriminate on the basis of race or gender, using the observable characteristics as proxies for the unobservable traits.[68] In Arrow's model, this discrimination becomes a self-fulfilling prophecy, as individuals in discriminated-against groups decide not to invest in their education because they will not be properly rewarded in the market.[69] Phelps and Arrow speak of observable traits and unobservable skills,[70] but their statistical discrimination model might well be adapted to observable traits and *impermissible traits*. It is often possible to make educated predictions about unknown and often-impermissible traits from the known traits available in the data.

Proxies for prohibited categories might be found in relatively innocuous activities such as web-browsing behavior. It is possible to predict race and other attributes with reasonable accuracy based on what pages an individual "likes" on Facebook.[71] According to one study using Facebook "likes,"

> African Americans and Caucasian Americans were correctly classified in 95% of cases, and males and females were correctly classified in 93% of cases, ... Christians and Muslims were correctly classified in 82% of cases, and similar results were achieved for Democrats and Republicans (85%). Sexual orientation was easier to distinguish among males (88%) than females (75%).[72]

The worry is that algorithms will utilize proxies for impermissible information, reconstructing with reasonable accuracy that barred information through analysis of available information.

Imagine a college admissions office turns over to an algorithm, whose mandate is to select students who will perform well in the job market. If the job market itself favors whites,[73] the algorithm may use available proxies for whiteness—such as the applicant's home zip code or the high school one attended—relying on the de facto segregation of housing and secondary education that has survived *Brown v. Board of Education*.[74] Viral discrimination

---

67. Kenneth J. Arrow, *The Theory of Discrimination*, *in* Discrimination and Labor Markets 3 (Orley Ashenfelter & Albert Rees eds., 1973); Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 Am. Econ. Rev. 659 (1972). Before Phelps and Arrow, others had postulated taste-based discrimination, where an employer had a preference for discrimination, which economists had suggested could not be maintained in a perfectly competitive environment. *See* Gary Becker, The Economics of Discrimination 84–85 (2d ed. 1971).

68. *See* Arrow, *supra* note 66, at 3–4.

69. *Id.* at 26–27.

70. *See generally* Arrow, *supra* note 66; Phelps, *supra* note 67.

71. Michal Kosinski et al., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 Proc. Nat'l Acad. of Sci. 5802, 5802 (2013).

72. *Id.* at 5803.

73. Barnes et al., *supra* note 18, at 988–89.

74. *See, e.g.*, John Iceland et al., U.S. Census Bureau, Racial and Ethnic Residential Segregation in the United States: 1980–2000 (2002), https://www.census.gov/prod/2002pubs/censr-3.pdf [https://perma.cc/X5RK-TQ7H].

could hold a pernicious quality absent in discrimination by human decisionmakers—the ability to defend the algorithm as a supposedly objective decisionmaker.

Thus, alongside this problem of intentional invidious discrimination, automated algorithms offer a perhaps more ubiquitous risk: replicating real-world inequalities. Discrimination can thus be propagated virally through real-world data. How should we address this problem? We now turn to this question.

III. Algorithmic Affirmative Action

The opposite of a black box society where "secret algorithms process[ ] inaccessible data"[75] is an "*intelligible* society."[76] Accordingly, Pasquale's principal solution to the problem of black box discrimination is transparency.[77] Pasquale uses "black box" in the following sense: "[A] system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other" (p. 3). Pasquale's metaphor of the "black box" suggests that the solution to algorithmic ills is to open algorithms up for examination. If the problem of algorithmic discrimination is likely to lie in manipulations, then indeed peering inside the black box seems the answer.

But if the problem is likely to lie in the ways that algorithms might replicate real-world discrimination through their statistical methodologies, then I suggest that the solution lies elsewhere. Instead of *transparency in the design* of the algorithm, what we need is a *transparency of inputs and outputs*. Pasquale himself recognizes that algorithmic ills often lie in the data, rather than the algorithm itself, but the "black box metaphor" is easy to misread (p. 18). By focusing on inputs and outputs, we can more readily identify disparate impact. Bertrand and Mullainathan's experiment[78] serves as an exemplar of this approach. A key question from the algorithmic affirmative action approach would be: Are African Americans or women (or another relevant group) receiving statistically worse results, given their relevant characteristics? The focus on outcomes rather than how an algorithm operates seems especially useful as algorithms become increasingly complicated, even able to modify themselves.[79]

---

75. *See* p. 14.

76. *See* p. 202.

77. Pasquale does not seek in this book to specify exactly what must be disclosed, by whom, to whom, or when. On all these details, he hints at flexibility, writing on the last page as his call to action: "[I]t is time for us as citizens to demand that important decisions about our financial and communication infrastructures be made intelligible, soon, to independent reviewers—and that, over the years and the decades to come, they be made part of a pubic record available to us all." P. 218.

78. *See supra* note 19 and accompanying text.

79. Nir Ailon et al., *Self-Improving Algorithms*, 40 Soc'y for Indus. & Applied Mathematics J. on Computing 350 (2011).

Transparency of the algorithms themselves can prove a limited solution. First, transparency invites manipulations by those who game those algorithms.[80] Google responds to those calling for algorithmic transparency by noting that while its page-ranking algorithm is described in general detail in public filings, "[i]f people who are trying to game search rankings knew every single detail about how we rank sites, it would be easier for them to 'spam' our results with pages that are not relevant and are frustrating to users—including porn and malware sites."[81] Moreover, Cynthia Dwork and Deirdre Mulligan argue that "[e]xposing the datasets and algorithms of big data analysis to scrutiny—transparency solutions—may improve individual comprehension, but given the independent (sometimes intended) complexity of algorithms, it is unreasonable to expect transparency alone to root out bias."[82] Second, requiring the publication of the algorithm itself may compromise trade secrets.[83] Third, individuals may know what the algorithm does yet lack choice in whether to participate in it.[84] Fourth, the algorithm may be too complicated for many others to understand, or even if it is understandable, too demanding, timewise, to comprehend fully.[85] Fifth, because the discrimination may arise through the training or operational data rather than the algorithm itself, revealing the facially neutral algorithm may help defend that algorithm from accusations of discrimination.[86] Finally, in the era of self-enhancing algorithms, the algorithm's human designers may not fully understand their own creation: even Google engineers may no longer understand what some of their algorithms do.[87]

This is why affirmative action is the right model for fashioning a remedy for algorithmic discrimination.[88] Here, I mean "affirmative action" in its

80. Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. Pa. L. Rev. 633, 639 (2017) ("The process for deciding which tax returns to audit, or whom to pull aside for secondary security screening at the airport, may need to be partly opaque to prevent tax cheats or terrorists from gaming the system." (manuscript at 6)).

81. Matt Cutts, *Google, Transparency, and Our Not-So-Secret Formula*, Google Europe Blog (Mar. 2, 2010), http://googlepolicyeurope.blogspot.com/2010/03/google-transparency-and-our-not-so.html [http://perma.cc/LBV8-QXHE].

82. Cynthia Dwork & Deirdre K. Mulligan, Response, *It's Not Privacy, and It's Not Fair*, 66 Stan. L. Rev. Online 35, 37 (2013), https://review.law.stanford.edu/wp-content/uploads/sites/3/2016/08/DworkMullliganSLR.pdf [https://perma.cc/5KF6-24SL].

83. Kroll et al., *supra* note 80, at 658.

84. *See* Dwork & Mulligan, *supra* note 82, at 37–38.

85. *See* Pasquale, *supra* note 8.

86. *See* discussion *supra* Part II.

87. Barry Schwartz, *Google's Paul Haahr: We Don't Fully Understand RankBrain*, Search Engine Roundtable (Mar. 8, 2016, 7:55 AM), https://www.seroundtable.com/google-dont-understand-rankbrain-21744.html [https://perma.cc/H7JK-NM5Q] ("RankBrain is Google's query interpretation that uses artificial intelligence.").

88. Cynthia Dwork and her colleagues suggest "fair affirmative action" as a goal of algorithmic decisionmaking, defining the phrase as obtaining "*statistical parity* (i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population), while treating similar individuals as similarly as possible." Dwork et al., *supra* note 62, at 214.

broadest sense, as a set of proactive practices that recognize deficiencies in the equality of opportunity and act in a multiplicity of ways to seek to correct for those deficiencies. The Affirmative Action Review conducted during the Clinton Administration defined affirmative action as "any effort taken to expand opportunity for women or racial, ethnic and national origin minorities by using membership in those groups that have been subject to discrimination as a consideration."[89] Affirmative action does not focus on identifying the *how* of discrimination, but on working to correct it, *regardless of its source*. For example, it does not ask if the hiring officer has biases unknown even to himself or herself, or whether structural reasons limit the number of applicants from a particular group. The EEOC's guidelines for affirmative action issued in 1979 cited Congress's finding of the "'complex and pervasive nature' of systemic discrimination against women and minorities" without worrying about identifying the sources of the discrimination precisely.[90] The goal is not to point fingers to the source of the problem, complex as it is likely to be, but to seek to rectify the problem.

The counterintuitive result of affirmative action is that the decisionmaker must take race and gender into account in order to ensure the fairness of the result. This is what struck Chief Justice John Roberts as implausible: "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."[91] The obvious remedy to the problem of manipulations of algorithms that produce racist or sexist outcomes would seem to be to mandate race or gender neutrality. In reality, however, even while neutrality is certainly better than hard-coded racism or sexism, racial or sex neutrality would in fact perpetuate the problem of algorithmic replication of existing racism. Justice Sonia Sotomayor responded sharply to Chief Justice Roberts' claim in a recent opinion: "The way to stop discrimination on the basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes open to the unfortunate effects of centuries of racial discrimination."[92] In the past, opponents of affirmative action have sought to prevent the government from collecting statistics on race, which would make it more difficult to establish wrongful discrimination and also make affirmative action more difficult. For example, California's proposed Racial Privacy Initiative (more accurately a Racial Blindness Initiative) would have foreclosed the collection of racial information by the

89. George Stephanopoulos & Christopher Edley, Jr., Affirmative Action Review: Report to the President § 1.1 n.1 (1995), https://clinton2.nara.gov/WH/EOP/OP/html/aa/aa-index.html [https://perma.cc/374L-32BG]. *See generally* John Valery White, *What is Affirmative Action?*, 78 Tul. L. Rev. 2117 (2004).

90. Affirmative Action Appropriate Under Title VII of the Civil Rights Act of 1964, as Amended, 44 Fed. Reg. 4422 (Jan. 19, 1979) (to be codified at 29 C.F.R. pt. 1608) (quoting H.R. Rep. No. 92-238, 92nd Cong., 2nd Sess. 8 (1972)).

91. Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 748 (2007).

92. Schuette v. Coal. to Defend Affirmative Action, Integration & Immigrant Rights & Fight for Equal. By Any Means Necessary (BAMN), 134 S. Ct. 1623, 1676 (2014) (Sotomayor, J., dissenting).

government, resulting in adverse effects on minorities in the state.[93] As I have written elsewhere,

> [The] initiative [to deny government racial information collection] would force racial indifference on state government by preventing it from gathering racial information—in employment, in education, and in law enforcement. Of course, government workers could secretly still make personal assessments of race—based on appearance, speech, domicile, and name. Rather, the measure would find its significance in denying the government any systematic ability to engage in affirmative action. At the same time, it would deny minorities the factual record that would facilitate claims for police harassment and governmental redlining.[94]

Corporations may be reluctant to explicitly consider protected categories in their decisionmaking for fear that this might be used to argue that they intended, or at least abided, any discrimination that persists. But such efforts to avoid and ameliorate discrimination should be recognized as exculpatory, not incriminating.

What would algorithmic affirmative action consist of? It would begin by recognizing the differential results possible even with ostensibly neutral algorithms. At times, this might mean different design choices, such as what information the algorithm provides at what time. According to the Federal Trade Commission, one company decided to exclude where a person lived "from its hiring algorithm because of concerns about racial discrimination, particularly since different neighborhoods can have different racial compositions."[95]

Consider the well-known problem that African Americans have faced in hailing taxicabs, an obstacle that can severely restrict mobility.[96] Uber designed its platform so that its drivers do not see a photo of the passenger until after the driver has accepted the fare. Moreover, Uber's platform does not give the driver the passenger's destination again until the driver has accepted the fare, preventing the use of the redundant encoding of home address as a proxy for race. A driver may cancel a fare when he or she learns of the passenger's race or destination, but every canceled trip is seen as a negative mark against the driver. The end result is that some African Americans report that it is easier to obtain transportation using Uber than through

93. James Q. Wilson, *Colorblind Versus Blindfolded*, L.A. Times (July 21, 2003), http://articles.latimes.com/2003/jul/31/opinion/oe-wilson31 [https://perma.cc/398J-CAYM].

94. Anupam Chander, Essay, *Minorities, Shareholder and Otherwise*, 113 Yale L.J. 119, 173 (2003).

95. Fed. Trade Comm'n, Big Data: A Tool for Inclusion or Exclusion, at v (2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf [https://perma.cc/C7MR-WW9S].

96. ABC news program Good Morning America demonstrated the difficulty that African American attorney Christopher Darden, who was the prosecutor in the O.J. Simpson case, faced when trying to hail a cab in the evening. Dan Harris & Gitika Ahuja, *Race for a Cab: When Hailing a Ride Isn't So Black and White*, ABC News (Apr. 1, 2009), http://abcnews.go.com/GMA/race-cab-hailing-ride-black-white/story?id=7223511 [https://perma.cc/N2GT-2JXR].

hailing a taxi on the street.[97] This is certainly no panacea to the problems of racism such as those experienced by the driver. For example, Nancy Leong has argued that racism will likely evince itself in the ratings system, leading black drivers to be less well-rated than white drivers.[98]

At times, a likely discriminatory result may make an algorithmic approach unwise. In response to the revelations of sexist results from its autocomplete for "women should," Google simply directed its computers not to perform autocomplete for those words (or for "men should").[99] For a time, searches for "n***** house" during Barack Obama's presidency would report the White House as a top search result.[100] Whether because of Google's response to these reports, or because of changes in users' behavior, that result no longer materializes.

An affirmative action approach recognizing viral discrimination would require a focus on the data that algorithms use. Transparency in data, rather than in the algorithms themselves, is consistent with some recent work in computer science. Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian seek to measure algorithmic fairness without needing transparency in the design of the algorithm: "Instead of requiring access to the [algorithm], we propose making inferences based on the data [the algorithm] uses."[101] This recalls Pasquale's suggestion of transparency in the data that algorithms operate upon: "When a company builds a dossier on you, you deserve a chance to review it and correct it," he writes (p. 147). And Pasquale suggests that a company not attending to the discriminatory results arising out of discriminatory data might be committing "algorithmic negligence" (p. 40).

Data about outputs would be crucial to identifying an algorithm's disparate impact. Richard Primus observes that disparate impact doctrine

97. Journalist Latoya Peterson writes that she, as an African American woman, turns to Uber to find transportation because regular taxis ignore her. Latoya Peterson, *Cab Drivers, Uber, and the Costs of Racism*, Racialicious (Nov. 28, 2012), https://web.archive.org/web/20160326193424/http://www.racialicious.com/2012/11/28/cab-drivers-uber-and-the-costs-of-racism/ [https://perma.cc/L24B-WXM4] ("[T]he premium car service removes the racism factor when you need a ride."); *see also* Clinton Yates, *Uber: When Cabs Whiz By, It's a Pick Me Up*, Wash. Post (Sept. 28, 2012), https://www.washingtonpost.com/blogs/therootdc/post/uber-when-cabs-whiz-by-its-a-pick-me-up/2012/09/28/06a41f0c-082f-11e2-858a-5311df86ab04_blog.html?hpid=z4 [https://perma.cc/W29B-C43W].

98. *Uber, Privacy, and Discrimination*, Nancy Leong (Apr. 20, 2014), http://www.nancyleong.com/race-2/uber-privacy-discrimination/ [https://perma.cc/9FDB-WSKY]; *see also* Benjamin Sachs, *Uber: A Platform for Discrimination*, On Labor (Oct. 22, 2015), https://onlabor.org/2015/10/22/uber-a-platform-for-discrimination/ [https://perma.cc/DT2S-9AP8].

99. *See* David Auerbach, *Filling the Void*, Slate (Nov. 19, 2013, 11:58 AM), http://www.slate.com/articles/technology/bitwise/2013/11/google_autocomplete_the_results_aren_t_always_what_you_think_they_are.html [https://perma.cc/3WAU-S4TG].

100. Samuel Gibbs, *Google Says Sorry Over Racist Google Maps White House Search Results*, Guardian (May 20, 2015, 5:52 AM), http://www.theguardian.com/technology/2015/may/20/google-apologises-racist-google-maps-white-house-search-result [https://perma.cc/DBB9-5TE3].

101. Michael Feldman et al., Certifying and Removing Disparate Impact 1 (2015) (emphasis omitted), https://arxiv.org/pdf/1412.3756.pdf [https://perma.cc/78LS-545L].

counters historical disadvantage: "By forcing employers to notice racial patterns and think about possible ways to change them, disparate impact doctrine helps diminish the power of historical hierarchies."[102] In the traditional application of disparate impact doctrine, employers were free to institute affirmative action programs based on statistical evidence that the percentage of minorities or women fell short of their percentage in the relevant pool.[103] Most recently, the Supreme Court has approved the use of statistics demonstrating disparate impact to support a claim for unfair housing.[104]

An affirmative action approach would seek to ensure that the data used to train an algorithm are evaluated for being embedded with viral discrimination. It would require attention to discrimination in both the validation set and the unseen test set of data.[105] The Obama Administration offered a similar approach in its recent report on big data: "To avoid exacerbating biases by encoding them into technological systems, we need to develop a principle of 'equal opportunity by design'—designing data systems that promote fairness and safeguard against discrimination from the first step of the engineering process and continuing throughout their lifespan."[106] Such an approach would require companies to anticipate how their algorithms are likely to operate in the real world and to review those operations for discriminatory results.

At times, it may be appropriate to share details of inputs and outputs with a third party who could review the fairness of the decisionmaker's algorithm. Computer scientists seeking algorithmic fairness have sometimes postulated both a decisionmaker, who uses the operational algorithm, and an independent certifier, who reviews the fairness of the decisionmaker's

---

102. Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 Harv. L. Rev. 493, 535 (2003).

103. As the Supreme Court announced in 1987:

> [C]onsideration of the sex of applicants for Skilled Craft jobs [may be] justified by the existence of a "manifest imbalance" that reflected underrepresentation of women in "traditionally segregated job categories." In determining whether an imbalance exists that would justify taking sex or race into account, a comparison of the percentage of minorities or women in the employer's work force with the percentage in the area labor market or general population is appropriate in analyzing jobs that require no special expertise, ... Where a job requires special training, however, the comparison should be with those in the labor force who possess the relevant qualifications.

Johnson v. Transp. Agency, 480 U.S. 616, 631–32 (1987) (citations omitted).

104. *See* Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2513 (2015).

105. Trevor Hastie et al., The Elements of Statistical Learning: Data Mining, Inference, and Prediction 222 (2d ed. 2009) ("If we are in a data-rich situation, the best approach for [model selection and model assessment] is to randomly divide the dataset into three parts: a training set, a validation set, and a test set. The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model.").

106. Exec. Office of the President, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights 5–6 (2016), https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf [https://perma.cc/9XQT-9VYQ].

algorithm.[107] In a recent blog post, Pasquale suggests that "[e]ven if algorithms at the heart of these processes 'transcend all understanding,' we can inspect the inputs (data) that go into them, restrict the contexts in which they are used, and demand outputs that avoid disparate impacts."[108]

Conclusion

When we appraise emerging technologies, we must be careful not to romanticize a pretechnological past. New technologies must be examined both in comparison to their less-technological alternatives and in the context of the world that we now inhabit. The black boxes of the past may have been analog, but they were every bit as obscure as the digital black boxes of today. They took place in committees, in conversations among executives, in backroom deals among power brokers, or most often, in the minds of men. In Lin-Manuel Miranda's musical celebrating the life of Alexander Hamilton, his rival Aaron Burr longs to "be in the room where it happens," where "the sausage gets made."[109] The ultimate black box, of course, is the brain which, even with the latest techniques, remains remarkably opaque.

The turn to algorithmic decisionmaking does not break us free from prejudices. This is one of Pasquale's most important contributions: the recognition that automated systems are not free of bias simply because they are executed by logical machines. Consider yet another recent example: "Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk."[110] Avoiding viral discrimination will require the application of affirmative action principles. Pasquale forces us to confront the ethics of the coming age of automated algorithms.[111] As Pasquale crucially observes, "[O]nly humans can perform the critical function of making sure that, as our social relations become ever more automated, domination and discrimination aren't built invisibly into their code" (p. 213).

---

107. Feldman et al., *supra* note 101, at 6 ("It is Bob's job to verify that on the data *D*, Alice's algorithm *A* is not liable for a claim of disparate impact.").

108. Pasquale, *supra* note 8.

109. Lin-Manuel Miranda, *The Room Where It Happens*, *in* Lin-Manuel Miranda & Jeremy McCarter, Hamilton: The Revolution 186, 187 (2016).

110. Daniel Victor, *Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk.*, N.Y. Times (Mar. 24, 2016), http://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html (on file with *Michigan Law Review*).

111. *See* Mustafa Suleyman, *Announcing the Partnership on AI to Benefit People & Society*, DeepMind (Sept. 28, 2016), https://deepmind.com/blog/announcing-partnership-ai-benefit-people-society/ [https://perma.cc/GFY7-C3EC]; *Industry Leaders Establish Partnership on AI Best Practices*, IBM (Sept. 28, 2016), http://www-03.ibm.com/press/us/en/pressrelease/50668.wss [https://perma.cc/SV65-U8TV].