APPS IN BLACK AND WHITE

Ruha BENJAMIN, *Race after Technology* (Cambridge UK, Polity Press, 2019, 178 p.)

Many large hospitals and insurers use complex computational instruments to model risk-adjusted patient outcomes at the cohort and individual patient level. About 200 million people a year are evaluated and ranked by such tools each year. Their outputs determine, among other things, whether a patient will be triaged to a hospital's "high-risk management program." This is intended to channel resources and provider attention to those patients who are at highest medical risk. In October 2019, a study was published in *Science* describing a high-risk management program in use at an unnamed large academic center in the United States. The study's central finding can be summarized in a single snapshot statistic: African-American and non-black patients assigned identical risk-scores by the predictive tool experienced statistically and materially different outcomes. Black patients had roughly 25% more chronic illnesses than similarly ranked patients who were not black.[1] The prediction thus had a greater rate of false negatives for African-Americans than for similarly situated others. The study further found that this disparity could be narrowed by eschewing historical cost as a proxy for the severity of illness and adopting an alternative outcome variable.

Should this predictive tool be condemned as a failure of racial equity, or, worse, as racist? These questions are, in my view, more complicated than commonly realized. There was no evidence to suggest that the designers of the algorithm had been spurred on by a specific intention to disadvantage racial minorities. Nor is the idea of predicting high-risk patients particularly new. Risk-adjusted models have been widely employed since the late 1990s as part of a hospital-accreditation process managed by the Joint Commission on Accreditation of Healthcare Organizations. Innovations in machine-learning over the last three decades have changed the way in which prediction is done, but the basic notion of using historical usage data to make predictions about future patient outcomes is a familiar one. Furthermore, the designers of the predictive

[1] Ziad OBERMEYER, Brian POWERS, Christine VOGELI and Sendhil MULLAINATHAN, 2019, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 366 (6464): 447-453.

Aziz Z. HUQ, University of Chicago Law School, Chicago, USA [huq@uchicago.edu]

European Journal of Sociology, 61, 3, (2021), pp. 423–431—0003-9756/21/0000-900$07.50per art + $0.10 per page
© *European Journal of Sociology* 2021. doi: 10.1017/S0003975620000223

tool considered in the *Science* study did not use patient race information. Racial disparities had to be identified subsequently by matching outcomes with self-reported race contained in patients' intake paperwork. A similar approach is used in other predictive tools that are in widespread use. The Compas instrument for pre-trial bail determinations across the United States, for example, does not use race as an input, and nevertheless has been interrogated on racial equity grounds. Racial data as an input, in any case, is an unreliable proxy for output disparities. It is mathematically possible for the omission of race information to generate predictions that are substantially less accurate for a minority group than for a majority. This can occur, for example, if a predictor relies on a trait that is highly correlated with relevant outcomes for the majority, but is only weakly correlated with those outcomes for a minority group. Race as a predictor of erroneous prediction, therefore, is not necessarily associated with race as a parameter of a predictive model. Perhaps the best way of describing the flaw in the hospital-based prediction tool is as a form of malign neglect embedded in the decision to use historical cost as a proxy for illness severity without considering the ways in which the two might diverge.

The *Science* study exemplifies a genre of scholarship and muck-raking journalist *exposés* of racial disparities arising from the private or state use of predictive technologies. Other studies have touched on welfare and unemployment management, bail and probation determinations, teacher hiring, and decisions by regulatory agencies of which firms to investigate. The sheer profusion of such interventions suggests that we are witnessing a collision: On one side are innovations in computational instruments of prediction and analysis; on the other is the complex of beliefs, dispositions, social practices and institutional arrangements known as "race." This collision has spurred an explosion of work in computer science seeking to define mathematically terms such as "discrimination" and "equality," and then to model formal corrections to mitigate discriminatory effects. The Fairness, Accountability, and Transparency in Machine Learning conference (or "FAT-ML"), first convened by Solon Baracas and Moritz Hardt in 2014, now attracts hundreds of researchers who are largely focused on technical questions of algorithmic design and accounting. Much of this literature attains clarity via a mathematical formalization of open-ended terms such as "discrimination" and "equality" coupled to a studied disregard of contextualizing detail. More glacial has been the advance of social scientific accounting of the same topics. Influential accounts by Cathy O'Neil, Virginia Eubanks, and Safiya Noble (among others) have illuminated specific pressure points

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press

in the predictive technology/race collision. But a comprehensive and definitive synoptic analysis has remained to date wanting.

Articulating a tractable analytic framework for interrogating the race/technology nexus is challenging because of the sheer volume of difficult technical, social, and normative questions it entails. Consider the puzzles posed just by the high-risk patent management program: What was the mechanism through which the social and symbolic category of race entered that instrument to generate a racial disparity in recommendations? Are there different pathways, in other social and institutional settings conducing to similar patternings? Would fixes create new costs, and if so how would those costs be distributed ? Even if it is self-evident that a prediction tool generates a higher rate of false negatives for blacks patients than for whites ones, and that this is morally problematic, are all racial disparities resulting from computational predictive tools equally troubling? Are disparities ever justified (for example, would disparities in the treatment of criminals justified if they mitigate disparities in victimization rates)? Do all disparate predictions—whether generated by search engines, recommender apps, risk-prediction instruments, or pattern recognition tools applied to visual or audio data—equally yield varying downstream effects across racial groups? And if a computational tool were to be replaced with a human decision-maker, would a different pattern of outcomes, whether for better or for worse, ensue? How, indeed, does the introduction of computational predictive tools alter social practices or individual beliefs relevant to race? For instance, does the availability of a computational risk management tool induce physicians to rely less on information they extract from patients—and if physicians are (all else being equal) more likely to err when dealing directly with black than white patients, might there be cross-cutting (and partly offsetting) effects fro substitution of human with machine decisions? Finally, congruent with the ordinary scope of sociological inquiry (although squarely within the bailiwick of legal scholars) lies the normative question of what corrections can or should be prescribed for the objectionable race-related effects of predictive technologies. A synoptic account would illuminate how these questions hang together; pick out the joints at which we should start unraveling various empirical, analytic, and normative questions; and even offer some lucidity on the vectors of influence to and from the social to the technological.

Ruha Benjamin's professed ambition in *Race after Technology* is to taxonomize comprehensively the ways in which new predictive technologies collide with race in the American context. Her book purports to be a "field guide into the world of biased bots, altruistic algorithms, and their

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press

many coded cousins" that "open[s] the Black box of coded inequity" [7, 34]. In reaction against an earlier generation of scholarship about the "digital divide," Benjamin discounts technology firms' and futurists' optimism about the positive effects of new digital tools. Celebratory or utopian aspirations for new digital technologies obscure how they work instead as channels through which ancient forms of racial subordination and marginalization are temporally extended or even expanded in reach. In Benjamin's analysis, race is both an input to computational predictions [59], and also "a tool of vision and division with often deadly results" [36]. This bilateral ebb and flow conduces to what she repeatedly denounces as a new "Jim Code" [5-6].

To anatomize the ensuing possibilities of race/technology interaction, she offers a quadripartite classificatory matrix. This comprises (1) engineered inequity, in which an instrument explicitly works to amplify racial hierarchies; (2) default discrimination, which entails the neglect of racial cleavages during an instrument's design; (3) coded exposure, in which a computer vision tool operates differently in respect to distinct phenotypes; and (4) technological beneficence, in which fixes for inequitable effects end up reproducing or deepening discriminatory processes [47-48]. These four categories map onto the four principal substantive chapters of the book. Their bleak perspective on technological change gives way in a closing chapter. Infused by (albeit without citing) the approach taken by prison abolition advocates such as Ruth Wilson Gilmore, Benjamin briefly gestures toward the possibility of predictive tools that distribute, rather than concentrate power. She imagines instruments that inverted the habitual regulatory focus upon racial minorities, and instead track non-minority white-collar criminality's costs [196].

There are, unfortunately, some obstacles to *Race after Technology* becoming a standard reference work for the taxonomy and analysis of race/technology interactions. An important initial concern is that its central terms—the eponymous pair of labels chief among them—go without gloss or definition. Its organizing taxonomy is also ambiguous and incomplete. Finally, the evidence tendered in the book does always not support the diagnosis or prescription that follows. As a work of analysis, therefore, it has some limitations. That, however, does not necessarily sap it of importance. For there are other ways of finding value in Benjamin's vigorously argued contribution.

Begin, though, with the analytic worries. At the threshold, Benjamin gives no definition of either of her key terms. "Technology" is a broad, contingent category, not a natural kind. Particular examples have divergent effects on racial logics. Many—think here of Eli Whitney's cotton

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press

gin—have worked to the considerable detriment of African-Americans. In contrast, the largest historical gains in African-American life expectancy occurred (during the Jim Crow dispensation) thanks to advances in urban water and sewage management—which is, again, a kind of technology, albeit somewhat less glamarous than big data tools.[2]

It isn't clear what bounds "technology" as Benjamin uses the word, and whether this term reaches both the cotton gin and the sewage system. Her text bristles with *au courant* terms such as "bots," "deep learning" and "social networks." But such references merely deepen the definitional puzzle. I assume, for example, that her argument does not extend to deep learning instruments applied to mammograms for the detection of breast cancer to the extent their save women's lives. On the other hand, her argument does seem to extend to race and gender discrimination in the "tech labor force," which is not obviously the result of any computational decision-making [58]. To the extent that Benjamin is read as grappling with what is colloquially known as the "tech" sector of the economy, moreover, there is no accounting in her text of products and services that have improved the lives of racial minorities.

Benjamin also uses terms such as "race," "racing," and "racism" without explanation or gloss. Except for a few fleeting mentions of Asian-Americans and Hispanic tech users, her "Jim Crow" rhetoric seems to frame her argument as a matter of African-American experiences with technology. Only fleetingly does the book glimpse the possibility that the technologies in question are global phenomena, which seed quite different dynamics of inclusion and exclusion in other socioeconomic contexts.

Yet more puzzling, though, is its silence in respect to how to understand the axial term "race." A familiar sociological approach is to take race as a "symbolic category" that is "misrecognized as a natural category."[3] But Benjamin is discussing the work of programmers and computer scientists who, as Ann Morning has shown, are likely still to hew to biological concepts of race.[4] It hence remains unclear whether and how the design or use of pertinent technologies rests upon assumptions of either race's biological or alternatively its symbolic character. As a result,

[2] Werner Troesken, 2015, *The Pox of Liberty: How the Constitution Left Americans Rich, Free, and Prone to Infection* (Chicago, IL, University of Chicago Press).

[3] Matthew Desmond and Mustapha Emirbayer, 2009, "What is racial domination?," *Du Bois Review: Social Science Research on Race*, 6 (2): 335-355.

[4] Ann Morning, 2011, *The Nature of Race: How Scientists Think and Teach About Racial Difference* (Berkeley, University of California Press).

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press

the precise conception of "race " at work with respect to specific technologies remains obscure.

Assuming that Benjamin herself adopts the disciplinary standard view of race as a symbolic category, it is still unclear how the deployment of predictive or communication technologies alter the symbolic content of racial terms. Contra the implication of her text, there is no reason to assume that digital technologies necessarily exacerbate the salience of phenotypical markers of identity, as opposed to facilitating social sorting and decision-making that does not hinge upon a racial term. As the legal scholar Lior Strahilevitz observed more than a decade ago, where race is a basis of statistical discrimination, a plausible strategy to mitigate discrimination is to enable individuals to offer more fine-grained, verifiable information that allows sorting on nonracial terms. As he predicted, laws that prohibit disclosures of criminal-history information have the perverse effect of inducing reliance on inferred race on the part of employers.[5] The path from digital technology to racial ideology, in other words, is not as stark or single-minded as Benjamin intimates.

Nor is a distictive pathway from technology to racial stratification more persuasively mapped out in the book. Let me offer one example to show this. A hoary chestnut in discussions of race and technology is the Compas bail algorithm used by criminal courts around the country. The Compas tool has been found to generate large racial disparities in rates of false positives. Citing this finding, Benjamin argues that Compas "builds upon" and "reinforces" "already existing forms of racial domination" [81]. Perhaps—but her thin analysis fails to demonstrate as much. The net effects on racial stratification of instruments such as Compas are, in fact, more complex than Benjamin allows for two reasons. First, a comparison of false positives rates between racial groups (which is the standard form of criticism) is only one way of measuring racial disparities, and not necessarily always the most important. Recall, for example, that what mattered in the hospital risk-management setting was the rate of false negatives. And it is not mathematically possible to eliminate all disparities between demographically distinct racial groups: Eliminating disparities in false positives, for example, yields other kinds of disparities, including in false negatives. Hence, the question in evaluating a predictive tool's outcome cannot simply be whether there are disparities (there always are), but whether these disparities matter in normative terms.

[5] Lior Jacob STRAHILEVITZ, 2008, "Privacy versus Antidiscrimination," *University of Chicago Law Review*, 75 (1): 363-381.

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press

Second, Benjamin incorrectly assumes in the bail context that African-Americans experience harms only from false positives (i.e., wrongful detentions). But when most crime is intraracial, a false negative (i.e., a failure to detain a dangerous person) will also impose racialized costs. Compas may well cause unjustified racial disparities.[6] But Benjamin's casual treatment does not illuminate their causes, magnitude, or cures.

Even setting to one side these definitional uncertainties, I am not sure that Benjamin's taxonomy provides a tractable framework for analysis or critique. Consider once more the hospital risk-management program with which this review began. I think this could be placed within her categories (1), (2), and (4). (That said, categories (1) and (2) are both described in terms of design choices that ramify in unequal ways, so I find them hard to tease apart more generally). That is, her taxonomy does not provide a means for sorting among examples for the simple reason that its categories are not mutually exclusive. Nor does it elucidate the cause of racial disparities. Locating the hospital risk-management program within Benjamin's framework provides no insight into *why* disparities emerge, or how to mitigate them. (Notice that simply switching to human decision-making might make the racial disparity worse, not better). As Benjamin duly notes, predictive instruments that rely on historical datasets—such as linguistic corpora, municipal crime data, or historical patterns of healthcare consumption—can generate predictions that ramify discriminatory practices: Those overpoliced or underserved with healthcare are likely to continue to experience undesirable, and even harmful, treatment in the future. But, as the *Science* study demonstrated, different choices about the kinds of historical training data upon which predictions are grounded will yield varying levels of racial disparity. Benjamin paints too coarsely when she characterizes these tools as simply "a mirror" to larger social realities, or a "part of the larger matrix of systemic racism" [77-78]. Perhaps sometimes, but not necessarily so. Sometimes, the disparity can be identified, and the predictive tool adjusted. Nothing Benjamin says helps elucidate *how* minorities obtain equal benefits from technological advances without shouldering disproportionate costs.

One reason for this difficulty may be a certain diffuseness in Benjamin's four central categories. At some points in reading the book, I had some difficulty identifying a clear argumentative thread uniting a given

[6] Aziz Z. Huq, 2019, "Racial equity in algorithmic criminal justice," *Duke Law Journal*, 68 (6): 1043-1134.

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press

chapter (and hence a given category). Take, by way of example, Chapter 4, although the exercise can be pursued with any one of her chapters. This is called "Technological Benevolence," even though Benjamin's introductory roadmap used the term "technological beneficence" [47, 137]. I assume nothing rests upon this terminological shuffle, except for an unnecessary nip of confusion for the reader. The chapter then skips from the effect of electronic monitoring on the carceral state's footprint; to hiring algorithms that use affect-recognition tools; to companies that aggregate and sell ethno-racial identification data by inferring race from residential ZIP codes; to the HBO show *Silicon Valley*'s mockery of techie jargon; to predictive instruments in healthcare settings that pick out (often minority) "super utilizers." This heterogeneous *tour d'horizon* proves for Benjamin that "those who genuinely seek social justice [should] avoid falling under the spell of techno-benevolence," and instead call for a "revolution" [158]. Really ? Even if Benjamin has demonstrated that putatively beneficial technological interventions can have unanticipated costs—not all her examples advance that point—it does not follow that all or most technological innovations have perverse racial consequences. Notice, rather, the uneasy echo in Benjamin's argument here of a famous Chicago School argument against safety regulation. Whereas the Chicago economists predicted that seatbelt regulation would necessarily induce more reckless driving (it didn't), Benjamin intimates that technological fixes will always be "duplicitous" [148]. In either case, a possibility theorem is not the same as a proof.

All this said, perhaps it misses the point to gloss *Race after Technology* in this fashion. Perhaps I have misconstrued Benjamin's purpose. I am writing for—and you are reading—a scholarly journal. But her text might be understood not as a scholar's perspective upon an important debate on social processes, and instead as a public intellecual's intervention into ongoing crystallizations of social meaning. Hence the evocative and soaring language; hence the use of categories that are verbally resonant if not analytically crisp; and hence the sheer breadth of critique. By conjuring an intellectual lineage to Michele Alexander's *The New Jim Crow*, Benjamin indeed invokes a different, and distinctly prophetic register. Much of her book, furthermore, inhabits an emotional space of (quite justified) anger and outrage against persisting social exclusion and injustice of a racial character. Her intervention, on this view, is not designed to isolate the causal effects of a given technology, or to specify in thick detail the operation of any particular predictive institution. It is rather to vocalize the awfulness of a wicked and pervasive "racism" that serves as guiding spirit to the technological world. Fair enough, this is a

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press

laudable and important enterprise. Indeed, if that is so, little that I have said here bears on whether *Race Against Technology* succeeds on its own terms—that is, whether it will inspire and move readers to salutary political action. On that score, I wish Benjamin every success.

AZIZ Z. HUQ

https://doi.org/10.1017/S0003975620000223 Published online by Cambridge University Press