Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

What's Wrong with Machine Bias

CLINTON CASTRO
Florida International University

Data-driven, decision-making technologies used in the justice system to inform decisions about bail, parole, and prison sentencing are biased against historically marginalized groups (Angwin, Larson, Mattu, & Kirchner 2016). But these technologies' judgments—which reproduce patterns of wrongful discrimination embedded in the historical datasets that they are trained on—are well-evidenced. This presents a puzzle: how can we account for the wrong these judgments engender without also indicting morally permissible statistical inferences about persons? I motivate this puzzle and attempt an answer.

1. Introduction[1]

In the spring of 2014 Sade Jones and Brisha Borden attempted to steal an unlocked Huffy bicycle and Razor scooter from the lawn of a Fort Lauderdale home. Before the two could flee the scene—and after they had realized the toys were too small for them—they were arrested and charged with burglary and petty theft. The items were valued at a total of $80.

Compare this crime with one that took place in North Lauderdale the previous summer. Vernon Prater, who had previously been convicted of multiple armed robberies, was arrested for stealing tools from Home Depot. The items were valued at a total of $86.

When Jones, Borden, and Prater were booked into jail, a computer program issued a prediction of the likelihood of each committing a future crime. The program—Northpointe Inc.'s Correctional Offender Management Profiling for

Contact: Clinton Castro <clinton.g.m.castro@gmail.com>

1. Information in this section pertaining to COMPAS and Jones's, Borden's and Prater's cases obtained from Angwin, Larson, Mattu, and Kirchner (2016).

https://doi.org/10.3998/ergo.12405314.0006.015

Alternative Sanctions, or COMPAS—issued some surprising predictions. This was Jones's first offense, and yet she was identified as medium risk. Borden—whose only priors were misdemeanors committed as a juvenile—was identified as high risk. And Prater—the seasoned criminal—came back low risk.

As it turns out, these predictions were wide of the mark. Prater has since been charged with 30 felony counts, including burglary and grand theft. He currently sits in Florida state prison for these crimes. Borden and Jones completed probation without re-offending.

The predictions issued by COMPAS weren't idle speculations. Its predictions, also known as *risk assessment scores*, inform decisions at various stages of the criminal justice system. In Borden and Jones's case, the scores arguably meant a higher bond amount. (From $0 to $1,000.)

It's natural to ask: what led COMPAS to issue such poor predictions in these cases? Unfortunately, we can only speculate. The future-criminal formula that COMPAS relies on is under proprietary lock and key.

Here's what we *do* know: these stories fit into a pattern. COMPAS falsely identifies black defendants as future criminals at almost twice the rate of white defendants.[2] It also falsely identifies white defendants as low risk more often than black defendants.

Jones and Borden are black, and Prater white.

Borden and Jones were wronged when they were falsely identified as high and medium risk to re-offend; but how so, exactly?[3] To answer this question, I will proceed as follows. In Section 2, I define some terms and locate COMPAS in its larger context. In Section 3, I canvass some of the literature on stereotyping and statistical discrimination, in search of an answer to our question. The search comes up short, but it's instructive. Among other things, it points in the direction of a promising answer to our question. In Section 4, I fill in some of the details to this answer. I argue that to make sense of the wrong that Borden and Jones suffered, we need an account that sets a variable threshold for the amount of information needed to make statistical inferences about persons, where the threshold is sensitive to practical factors (such as the harms and benefits associated with being misclassified). We also need a pluralistic account to make full sense of machine bias; not all biased judgments are wrong in the same way.

2. N.B., of the 137 data points that COMPAS uses in issuing predictions, race is not one of them. To be sure, many of the data points it tracks—e.g., "Was one of your parents ever sent to jail or prison?"—do correlate with race. See the full list of data points here: https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html.

3. For the purposes of this essay, I assume that Borden and Jones were wronged. My hope is that in answering the central question of this essay I provide grounds for thinking this assumption is true.

2. Machine Bias

Programs like COMPAS are increasingly popular in courtrooms around the nation (Angwin et al. 2016). This is part of a general trend in the increased use of data-driven decision-making:

> Advertisers want data to reach profitable consumers, medical professionals to find side effects of prescription drugs, supply-chain operators to optimize their delivery routes, police to determine where to focus resources, and social scientists to study human interactions. (Barocas & Selbst 2016: 673)

But data—as the above quoted authors note—"is not a panacea" (Barocas & Selbst 2016: 673). Indeed, there is serious concern—as one White House report put it—"that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace"(United States Executive Office of the President 2014). And corrections, we might add.

How might these eclipses happen? Programs like COMPAS generate judgments by finding patterns in large datasets and basing their predictions off of those patterns (Finlay 2014). If those patterns reflect an underlying pattern of wrongful discrimination, the program's future judgments may reflect—and thus recreate—those underlying patterns.

To give a sense of how this can happen, consider the following case. A firm wants to automate decisions relating to hiring. They have a large data set with information about their employees, which includes employee zip codes and length of tenure at the firm. The zip codes, we can imagine, correlate with race: most of the firm's past and present white employees live in one zip code, and most of the firm's past and present black employees live in another. Further, there is a strong anti-black bias in the employee review process, and for this reason turnover among black employees is high. The firm identifies long tenure as one element of being a good employee and mines their data set to find correlations between tenure and other data points. The firm discovers that zip code is a strong predictor of length of tenure and sets up a scoring system where applicants gain or lose points based on their zip code. Obviously, the automated system will discriminate against black people. Further, it inherited this discriminatory disposition from a pattern of discrimination that was present in its training data. We can imagine that the firm had the best of intentions in setting up this system. Perhaps they learned that they had a reputation of being biased against black people. In an effort to eliminate human bias from their decision processes, they automated as much decision making as possible. The problem, of course, is that

their automated system has inherited the bias of its human predecessors.[4]

Plausibly, this is what happened with COMPAS. Indeed, Northpointe Inc. published a paper defending COMPAS where they identified its biases as "a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores" (Northpointe Inc. 2016a: 8). In other words, it's okay that COMPAS falsely identifies black defendants as future criminals at a higher rate than white defendants; after all, black people are disproportionately likely to be arrested.

This is how things might look from Northpointe's perspective, but Borden and Jones are likely to have a different take. Namely, that COMPAS is problematic because it's biased against them. This brings us to an important question: what does it mean for a program like COMPAS to be biased?

There are two distinct conceptions of bias that will be relevant for our purposes here. One, call it the *parity* conception, understands a judgment generated by the application of some method (M),[5] that purports to determine whether individuals are P (where P just is some property), to be biased if M is more likely to falsely judge individuals with social identity S as P (or not-P) than it is members of some appropriate comparison group.[6] Parity requires that COMPAS has similar false negative and false positive rates for white and black defendants. The other conception, call it the *calibration* conception, identifies a judgement as biased if it was generated by the application of some method (M) and M's outcomes are not independent of one's social identity (Corbett-Davies & Goel 2018). Calibration requires that black defendants identified as high risk by COMPAS re-offend at the same rate as white defendants identified as high risk by COMPAS. I will understand a judgment to be an instance of *machine bias* if it is biased by the lights of parity or calibration, and the entity issuing the judgment is a computer program.

A complication that will be discussed later on is that there are some situations where parity and calibration can't both be satisfied.[7] Indeed, there are cases where one can only be satisfied at the cost of the other, and ours seems to be one of them.[8] That is, given disparate arrest rates between white and black people, a risk assessment score must choose between satisfying parity and satisfying calibration.

4. For further discussion of the ways machines can inherit bias, see Barocas and Selbst (2016: esp. 677–691).

5. What do I mean by method? I have in mind the underlying procedure or process—formal or informal—that produces the judgment.

6. For further discussion of parity, see Corbett-Davies and Goel (2018).

7. For discussion of this issue, see Corbett-Davies and Goel (2018).

8. To the extent that COMPAS is biased because it is impartial (by at least one definition) and that it seems inevitable that it is biased if impartial (if it satisfies parity it will be biased in another way), the challenge of criticizing COMPAS is an instance of what Antony (1993) has called the "bias paradox," criticizing impartiality because it is biased while leveraging impartiality as part of that criticism.

More specifically, it must choose between having higher false positives and lower false negatives for black defendants and systematically giving white defendants higher risk scores than equally risky black defendants.

By the lights of our definitions, COMPAS is biased. But this does not yet mean that using COMPAS is wrongful. Using a biased machine or heuristic might—as Northpointe Inc.'s defense of COMPAS indicates—be justifiable.

Before turning to our investigation of where the line between appropriate statistical discrimination and problematic statistical discrimination lies, let's ask: why the focus on *machine* bias? Isn't the problem with COMPAS engendered by its *bias*?

The main reason for focusing on machine bias is that it is presumably different from typical cases of bias in ways that are potentially morally relevant. Machines' biased judgments often have a higher degree of epistemic warrant than those of humans. After all, programs like COMPAS are the product of the application of rigorous statistical methods. So, indicting a machine is likely to be more difficult because machine bias—as opposed to human bias—is more likely to be defensible from an epistemic point of view. And, as we will see, focusing on these cases will force us to think about stereotyping and statistical discrimination in new ways. Further, machines are playing an ever-increasing role in decision making. And, so, there is some urgency in better understanding the moral dynamics of automated decision-making.

3. Stereotyping and Discrimination

Let us now return to our question: how might we account for the wrong Borden and Jones suffered? To approach our question, I'll begin by looking at some recent work on stereotyping and statistical discrimination.

3.1. First Pass: Treating Persons as Individuals

Lawrence Blum identifies the wrong of stereotyping as grounded in a failure to treat persons as individuals:

> being seen as an individual is an important form of acknowledgment of persons, failure of such acknowledgement is a moral fault and constitutes a bad of all stereotyping. (Blum 2004: 272–273)

It is not entirely clear what Blum thinks treating persons as individuals amounts to. For the purposes of this essay I will understand him as making two claims:

> **GROUP**: x fails to treat y as an individual if x's treatment of y is solely based on y's group membership, and

WRONG: it is wrong to fail to treat a person as an individual.

Something about these claims seems right. Our treatment of persons should have an eye on the content of *their* character,[9] not that of the group(s) they belong to. This important insight notwithstanding, Blum's account is inadequate for our purposes.

Many judgments about persons are based on characteristics of the groups to which they belong and yet are not wrong as such. Credit scores, for instance, are based on the various statistical groups one belongs to (e.g., people who paid their bills on time). According to Blum's claims, then, there's something necessarily wrong with credit scoring (and related practices, such as setting insurance rates). And, importantly, the wrongs involved in generating credit scores and insurance rates are of the same sort and presumably degree that Borden and Jones suffered. But this seems wrong.[10] So, there are important moral differences between cases of statistical discrimination that Blum's account is not alive to. Namely, sometimes statistical discrimination is wrong—such as in the case of Borden and Jones—but sometimes it is not—such as in cases of unbiased credit and insurance scoring, and an account of the wrong of stereotyping must track these differences. Additionally, credit scoring and the generation of insurance rates do not seem to be morally problematic as such.

Further, certain cases of stereotyping do not seem wrong. The present account cannot account for this. Consider Erin Beeghly's (2015) case of a panicked father in an emergency room:

> **Doctor.** "Where is the doctor?" he might yell, "we need a doctor." The man might grab the first person he sees in a white coat, relying on the stereotype that doctors wear white coats, not caring that he is grabbing this or that particular doctor, not caring about the doctor at all in his or her individuality.

"There is no moral wrong in failing to recognize individuality here," Beeghly notes, "no moral vice" (2015: 688).

For these reasons, we cannot rely on Blum's account to make sense of the wrong of machine bias. But there is an important insight in it that we shouldn't forget: failure to appreciate one's individuality is an important way in which

9. One might wonder if we can judge the content of one's character without making reference to the group they belong to. I think that the following example suggests that we can. If I want to know whether my ride will be on time, I could think back into my personal history with her: Was she on time last week? The week before? This seems to be a case where I can make a judgement about someone's reliability or promptness without relying on data about a group they belongs to.

10. This is not to say the industry is perfect; far from it. Credit scores, too, have issues with bias (Rice & Swesnik 2013).

stereotyping and certain forms of discrimination wrong individuals. To preserve this insight we need a better theorized account of what it means to treat persons as individuals.

3.2. Second Pass: Use All the Available Information

Kasper Lippert-Rasmussen (2011) offers the following improvement on GROUP:

> **INFO**: *x* treats *y* as an individual if, and only if, *x*'s treatment of *y* is informed by all relevant information reasonably available to *x*.

Combine this with WRONG, and we have an account of the wrong of stereotyping that addresses several of our concerns with GROUP.

INFO doesn't overgeneralize as badly as GROUP—it morally permits at least some judgments that are based solely on one's group membership. It also gives us the tools to say why certain forms of statistical discrimination are worse than others: cases of biased judgment where one's full information is not used are, per INFO, worse than those where all the information is used.

This account, however, is not without its own set of shortcomings.

Borden, Jones, and Prater's predictions were generated using the same 137-question questionnaire. The COMPAS questionnaire[11] is quite thorough. It asks about a defendant's current charges, criminal history, history of "non-compliance" (history of parole, probation, court appearances, and crimes committed while on pretrial release), family's history of criminality, peer group behavior (their criminal history, gang activity, and drug use), substance abuse, living situation, social environment (the general level of criminal activity in one's neighborhood), education, vocation, level of boredom, level of happiness, feelings of discouragement, level of social isolation, attitudes and dispositions regarding honesty, perception about how they are viewed by others, level of anger, and moral attitudes and dispositions (e.g., whether the defendant agrees that a hungry person has a right to steal). COMPAS's predictions are the result of using all of this information.

INFO would presumably have us treat the cases alike. But this can't be right. There seems to be an important asymmetry between Borden and Jones's and Prater's case. Borden and Jones seem to have been wronged in a different way or to a greater degree than Prater. To be sure, it is arguable that Prater was indeed wronged by his treatment; at the very least, it's plausible that he was not treated as an individual (which, as Doctor shows, does not always involve a wrong). That INFO is not alive to the asymmetry between Borden and Jones's and Prater's cases is a problem for that account. Further, it is arguable that Prater was not wronged by the treatment he was subjected to.

11. Available here: https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

INFO arguably predicts this last detail, but at a high cost. As stated above, the COMPAS survey is quite detailed. Presumably, it was designed to gather as much information relevant to its predictions as it reasonably could. But if this is the case, then INFO renders the verdict that none of our parties were wronged by COMPAS.[12] Insofar as we are interested in understanding the wrong Borden and Jones suffered, this is problematic.

Further, INFO overgeneralizes in cases of credit scoring and insurance. In the United States, the Equal Credit Opportunity Act (ECOA) forbids creditors from discriminating on the basis of race, color, religion, national origin, sex, marital status, or age. At least some of this information is available to credit agencies and indicative of creditworthiness. When credit agencies and insurers have this relevant information and refrain from using it, they—per INFO—fail to treat persons as individuals. But this seems wrong. To the extent that credit agencies and insurers ignore this information, they *respect* our individuality.

3.3. Third Pass: Respect for Autonomy

Both INFO and GROUP share in common a certain sort of rigidity: they give a one-size-fits-all constraint for how to incorporate one's evidence into one's judgments (never treat an individual solely based on group membership, in the case of GROUP; always use all of the relevant information reasonably available to you, in the case of INFO). This rigidity is at least part of why both accounts overgeneralize.

Benjamin Eidelson (2013) offers an account that has some flexibility built into it. Rather than give a rigid directive as to how to treat our evidence in all cases, Eidelson directs us to give reasonable weight to evidence of ways an individual has exercised her autonomy. This leaves open whether we must in a specific case make a decision based solely on group membership or use all the information at our disposal. This feature of his account helps with some of the difficulties that INFO and GROUP encountered. Here is the account:

> **AUTO**: x treats y as an individual if, and only if:

12. One may respond by saying that a program like COMPAS leaves out reasonably available information, at least in certain cases. The program, for example, does not take the defendant's statement at the court into account, and this is a least sometimes relevant to predicting the likelihood that one will re-offend. This worry about the use of COMPAS is legitimate, and INFO's insight does help us put a finger on it. However, this does not affect my broader argument against INFO. We can imagine cases—perhaps Borden's and Jones's are examples—where no such information is present (perhaps the statements they provided were unremarkable and not predictive of anything) and COMPAS, because it is biased, misclassifies a defendant as high risk. In such cases, INFO renders the verdict that the defendants are not wronged and, yet, the problem persists. This, I think, shows that INFO is deficient, even if it can in some cases say that COMPAS fails to treat certain defendants as individuals because it leaves out reasonably available information, such as a defendant's statements.

> *Character Condition*: *x* gives reasonable weight to evidence of the ways *y* has exercised her autonomy in giving shape to her life, where this evidence is reasonably available and relevant to the determination at hand, and
>
> *Agency Condition*: if *x*'s judgments concerns *y*'s choices, these judgments are not made in a way that disparages *y*'s capacity to make those choices as an autonomous agent.

Let us begin by teasing out the difference between the character condition and the agency condition.

Consider the following case.[13]

> **Ray**. An employer, Molly, discriminates against female applicants because they are more likely to take parental leave. One particular candidate who suffers this discrimination, Ray, has a history of putting her professional life before her personal life and deeply opposes the idea of having a family. There is even evidence of this in her file: she's a columnist who writes about the joys of being a professional and against having a family. Molly fails to incorporate any of these details into her judgment about whether to interview Ray; she hastily rejects Ray's file on the grounds that she is a woman.

Rather than consider the ways in which Ray has chosen to author her life, Molly focuses solely on statistical facts about a group that Ray falls into. The character condition is thus not satisfied in this case and Molly fails to treat Ray as an individual. The agency condition, on the other hand, is satisfied, but only vacuously. Molly doesn't make a judgment about whether Ray's choices are being made autonomously or not.

Let us now re-imagine the case such that the agency condition is violated. This time, Molly is a bit more scrupulous. She seriously considers Ray's file. However, she thinks that Ray is at an age where her commitments to her profession will soon be taken over by a biological impulse to be a mother. In this case, the agency condition is violated because Molly sees Ray as heteronomous; she sees Ray as simply being driven around by her first-order desires.

That the Character Condition asks that we give reasonable weight to evidence of the ways an agent has exercised her autonomy enables AUTO to succeed where the previous two accounts failed. Consider, again, the credit scoring case. AUTO is not committed to saying that it is wrong for the credit card companies to blind themselves to certain demographic information. Looking at one's past actions (e.g., their history of on-time payments), as opposed to their race and gender is a

13. This draws from Eidelson's (2013) own illustrative case.

reasonable means for paying appropriate attention to the ways in which one has shaped one's own life.[14]

AUTO's merits notwithstanding, it is not able to help us make sense of the case of machine bias that is the focal point of this paper. Recall that Prater had previously been convicted of multiple armed robberies but that COMPAS judged that he was low-risk. COMPAS's insensitivity to evidence of the ways Prater has chosen to exercise his autonomy seems to be best explained by the fact that it gives too much weight to demographic factors.[15] So, the Character Condition is not met in Prater's case. But if it isn't, then it isn't met in Jones's, and Borden's cases either. Now it looks like the presumably important asymmetry between the cases is lost once more. Further, it does not seem that Prater was wronged when he was falsely identified as low-risk, though it does perhaps seem that he was not treated as an individual.

Let us now turn to the Agency Condition to see if it can recover the asymmetry. Whether COMPAS disparages one's capacities to make choices as an autonomous agent is an interesting question. On the one hand, the fact that it seems to be sensitive to race is one reason to think that COMPAS views defendants as heteronomous, which is at odds with the Agency Condition. Indeed, many of the questions asked by COMPAS track factors beyond the agent's control, suggesting that COMPAS sees agents more like objects pushed around by external forces than as autonomous beings. It asks questions like, "Was your father [...] ever arrested [...]?" "Was your mother [...] ever arrested [...]?" "Were your brothers or sisters ever arrested [...]?" (Northpointe Inc. 2016b: 3). Yet, the questionnaire also asks about one's personal history of crime, education, and employment. And it further inquires about one's attitudes towards claims like, "A hungry person has a right to steal" (Northpointe Inc. 2016b: 8). Because the questionnaire considers one's own personal history and specifically inquires about one's judgments about moral matters, it is very difficult to determine whether COMPAS takes a view of persons that disparages their ability to make autonomous choices. Further, it seems that COMPAS either violates the condition in Prater's, Jones's, and Borden's cases or violates it in none of their cases. So, even if the Agency Condition is violated, it cannot recover the asymmetry we were looking for.

It is for these reasons that AUTO lacks the resources to help us understand our case.

---

14. Though, of course, it is not perfect by any means. Because of discrimination, members of certain groups will have troubles with housing, employment, and the law that members of other groups will not. These factors could negatively affect one's ability to, say, make bill payments on time, and for reasons that are not a reflection of one's character (because those choices were not freely made).

15. N.B., this also explains the machine's hair trigger for identifying black defendants as future criminals.

3.4. General Remarks

Let's briefly take stock. Blum (2004) provides us with a promising starting point for thinking about our case: it renders the verdict that Borden and Jones were wronged because they were not treated as individuals. This largely seems right. Indeed, when COMPAS was used in a different case to inform the length of Eric Loomis's prison sentence, he appealed (unsuccessfully) on the grounds that using COMPAS violated his constitutional right to an individualized sentence (*State v. Loomis* 2016).[16] But Blum's account falls short because it overgeneralizes: it deems all statistical discrimination wrongful; as Doctor demonstrates, this can't be right.

Lippert-Rasmussen provides a helpful way forward. His account offers an explanation of why some (but not all) cases of statistical discrimination are wrongful: statistical judgments are wrongful when they are not informed by all the information one has. Unfortunately, his account both proves too much and too little. First, COMPAS arguably used all of the information at its disposal, so we cannot use this account to explain why Borden and Jones were wronged. Further, credit agencies and insurers do not use all of the information at their disposal as a means of avoiding wrongful discrimination. This is not wrongful, but INFO judges it as such.

Eidelson offered an an account that seemed able to better navigate our growing list of problem cases. His account tells us to give reasonable weight to evidence of the ways one has exercised her autonomy in giving shape to her life. This aspect of the view allowed it to properly deal with the case of credit scoring. However, it failed to account for important asymmetries in the main case under consideration in this paper. On a plausible reading of Eidelson's account, everyone was wronged and to a roughly equal degree. This seems to be the wrong result.

One general lesson to be observed here is that it is difficult to give a general account of the wrong associated with well-informed, but biased, statistical judgments without also condemning statistical judgments that aren't problematic. With this thought in mind, let's turn to a more promising approach to this challenge.

4. What's Wrong with Machine Bias

4.1. Fourth Pass: Risk of Misclassification

The accounts we have considered focus on epistemic considerations. GROUP asks: is the treatment based purely in statistical inferences? INFO asks: does it make full use of the available evidence? AUTO asks: does it give reasonable weight to

16. The Wisconsin Supreme Court responded by stating that the risk assessment score provided by COMPAS was not the sole basis for the sentencing decision.

evidence of the ways one has exercised her autonomy?

Perhaps our account should tether the amount of information one must use to some practical factors, such as the harms associated with misclassification. Here's a first pass at such an account:

> **RISK**: x fails to treat y as an individual if, and only if, x's treatment of y exposes y to a risk of being misclassified that y could reasonably reject.

I will understand the risk of being miscategorized o to an agent y, $R_y(o)$, as a function of the probability that o obtains, $Pr(o)$, and cost o imposes on y, $C_y(o)$, such that $R_y(o) = Pr(o) \cdot C_y(o)$. As far as reasonableness is concerned, we will work with a pretheoretical understanding.

On this account, then, there are two values to keep an eye on: the risk of being miscategorized and a threshold of acceptable risk, let's call it $T_y(o)$, which is set by what's reasonable for y to accept. We have defined the sorts of things that determine one's level of risk, the chances of being miscategorized and cost of being miscategorized to y. And, we have at least an intuitive sense of the sorts of things that can move the threshold: the stakes involved, considerations of fairness, the availability of less risky alternatives, and so on.

Let us, then, discuss Borden, Jones, and Prater's cases with an eye on the factors that affect these two values.

Begin with risk. In Borden and Jones's case, there are two factors pushing $R_{Borden}(o)$ and $R_{Jones}(o)$ up. First, misclassification is associated with a high cost—being identified as future criminals meant posting a bail of $1,000, each. Another factor pushing $R_{Borden}(o)$ and $R_{Jones}(o)$ up is the increased probability of misclassification that they face in virtue of being black. We cannot say exactly what the probabilities are here, but they are elevated relative to the general population. The disparity is even greater with respect to white people (Angwin et al. 2016).

Now consider Prater. For him, there is no cost to speak of. Because he is white, his chances of being misclassified as low risk are elevated. But this is negligible, since being misclassified—for him—is presumably costless. So, even though we do not know the exact probabilities, we can say the following: $R_{Borden}(o) > R_{Prater}(o)$ and $R_{Jones}(o) > R_{Prater}(o)$.

Let's now take a look at thresholds for acceptable risk. I will here focus on two factors relevant to T: stakes and considerations of fairness.

Begin with stakes. Borden and Jones share in common a reason for $T(o)$ to be fairly low. The costs of being identified as higher risk—longer prison sentences, more probation, and so on—are quite high. This puts pressure on COMPAS to be accurate.

Now turn to fairness. There are two considerations worth discussing on this front. The first is the fact that COMPAS treats two salient social groups—white people and black people—differently, in a domain where they deserve equal

treatment. That is, the brute fact that COMPAS is biased is itself a good reason for members of the group that it is biased against to be intolerant of the degree of risk it exposes them to.

These considerations alone may give us sufficient grounds to argue that Borden and Jones are wronged when they are subjected to COMPAS. However, it may not be enough to give us a general understanding of the wrong of machine bias. The second consideration of fairness, to which we will now turn, has not to do with fairness *within* the system under consideration. Rather, it looks to the larger context within which that system is situated.

When a program like COMPAS is constructed, we know that it will not be infallible. This raises a question of how to distribute its errors. Should we treat parity as a constraint? Or, should we opt for calibration? When one is working with the historical data that COMPAS is trained on, we can't have both. If one opts for calibration, the result is a tendency to disproportionately identify members of a certain population—that is, black people—as high risk. This is what Northpointe has in mind when they say COMPAS's bias is "a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores" (Northpointe Inc. 2016a). For Northpointe, calibration is the true definition of fairness. This line of reasoning may be able to justify at least some biases. And, in fact, it seems to offer at least some justification for tolerating the bias associated with COMPAS. But abiding calibration imposes a cost on a certain group. Whether the imposition of that cost on that group is justified, all things considered, depends on a number of factors. One of those, as I will now argue, is whether members of that group are disproportionately mistreated in other facets of their lives.

As the above discussion illustrates, a choice to satisfy calibration in this domain is a choice to disproportionately falsely identify black defendants as future criminals. The problem with this, in turn, is that it's already disproportionately costly to be black in America. Black people face harmful discrimination in housing (Turner, Santos, Levy, Wissoker, Aranda, Pitingolo, & The Urban Institute 2012), employment (Quillian, Pager, Hexel, & Midtbøen 2017), access to credit (Ghent, Hernández-Murillo, & Owyang 2014), and elsewhere. Black people are also subject to any number of abuses that aren't so easily quantified. As President Obama remarked in response to the George Zimmerman verdict:

> There are very few African-American men in this country who haven't had the experience of being followed when they were shopping in a department store. That includes me.
>
> And there are very few African-American men who haven't had the experience of walking across the street and hearing the locks click on the doors of cars. That happens to me, at least before I was a senator.

> There are very few African-Americans who haven't had the experience of getting on an elevator and a woman clutching her purse nervously and holding her breath until she had a chance to get off. That happens often. (Obama 2013)

It is against this backdrop that COMPAS imposes yet another cost on black people. These considerations, for black defendants, drive $T(o)$ down as well. In a society where one group (in this case, black people) continuously has costs disproportionately imposed on its members, when the society is forced to make a choice that will inevitably impose a cost on one group or another (in this case the choices seem to be black defendants, white defendants, and the public at large) members of the group seem to come to the bargaining table with a prima facie reason, grounded in fairness, not to have the cost imposed on them, and this is in virtue of the fact that they disproportionately shoulder the cost of a variety of other social arrangements.

Before moving on, it is worth pausing to discuss a complication that the trade-off between calibration and parity introduces. Specifically, what does the fact that we must choose between calibration and parity mean for what has been said so far? If this fact interacts with anything that has been said so far, it's the observation that in virtue of COMPAS's treating white and black defendants differently, black defendants have a reason to lower their threshold. The response to this on behalf of Northpointe is that COMPAS treats the two parties differently for good reason: it's the fair thing to do because the only other option, that is, simply sacrificing calibration for parity, is patently unfair. This response is inadequate—simply sacrificing calibration for parity is not the only other option (nor is it obvious that it is the wrong thing to do). As Corbett-Davies, Pierson, Feller, Goel, and Huq observe,

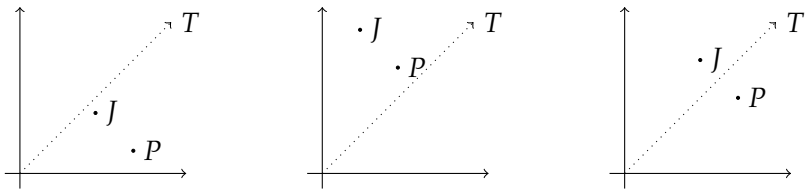
> [B]y collecting more data and accordingly increasing the accuracy of risk estimates, one can lower error rates. Further, one could raise the threshold for detaining defendants, reducing the number of people erroneously detained from all race groups. Finally, one could change the decision such that classification errors are less costly. For example, rather than being held in jail, risky defendants might be required to participate in community supervision programs. (2017: 8)

These, along with the question of whether to go in for calibration or parity, belong in the space of relevant alternatives. I do not know which of these (or which combination of them) should be pursued. But that is besides the point. The main point here is that we cannot say that COMPAS is justifiable as it stands because our only other option is to sacrifice calibration. We could instead, for example, raise the threshold for detaining defendants *and* sacrifice calibration for parity. Or, if

we think calibration and parity are both important because of the stakes involved, we can rethink the cash bail system altogether.

Because COMPAS is biased against black people (and not white people) *and* because black people face bias in other facets of their life that white people do not, $T_{Borden}(o) < T_{Prater}(o)$ and $T_{Jones}(o) < T_{Prater}(o)$.

We are now in position to make at least one significant remark about the case we began with. One problem other accounts had was making good on the intuitive idea that Borden and Jones received worse treatment than Prater. We can now offer an underwriting of that intuition. To see this, let's compare Prater and Jones's cases.[17] From what I have argued, it follows that $T_{Jones}(o) < T_{Prater}(o)$ and $R_{Jones}(o) > R_{Prater}(o)$. This leaves three possibilities for how the risk levels and thresholds relate to one another: (1) both Jones and Prater are imposed a reasonable cost (i.e., for both of them $T(o) > R(o)$); (2) neither Jones nor Prater are imposed a reasonable cost (i.e., for both of them $T(o) < R(o)$); (3) Prater is imposed a reasonable cost but Jones is not (i.e., $T_{Prater}(o) > R_{Prater}(o)$, but $T_{Jones}(o) < R_{Jones}(o)$). Represented visually:[18]



(1) (2) (3)

In all of these outcomes, Jones's treatment is worse than Prater's because, given the inequalities, she either receives acceptable treatment but has a smaller gap between the cost imposed on her and her threshold, is—like Prater—treated unacceptably but has a larger margin between the threshold and risk imposed on her, or, finally, is—unlike Prater—treated unacceptably.

Given our previous discussion of the costs to each party, it is arguably the case that $R_{Borden}(o)$ exceeds $T_{Borden}(o)$, $R_{Jones}(o)$ exceeds $T_{Jones}(o)$, and that $R_{Prater}(o)$ does not exceed $T_{Prater}(o)$. That is, we seem to be in situation (3).

So, RISK is able to explain the intuition that we started with: Borden and Jones were wronged and Prater was not.

17. Everything I draw from the comparison of Prater and Jones also applies to the comparison of Borden and Prater; the choice of using Jones was arbitrary and merely for the purposes of illustration.

18. In these graphs, J stands for the risk imposed on Jones ($R_{Jones}(o)$), P for the risk imposed on Prater ($R_{Prater}(o)$), and T for the threshold of acceptable risk. The exact placing of J and P are mostly arbitrary, as is the slope of T. What is not arbitrary is that T is at an incline, that J is lower on the x axis than P, and that J is higher on the y axis than P.

4.2. RISK and Problem Cases

4.2.1. Old Problem Cases

As we saw above, it is difficult to render the verdict that certain well-informed, but biased, statistical judgments are wrong without also condemning statistical judgments that aren't problematic. Let's return to the cases of innocuous statistical judgments that caused problems for GROUP, INFO, and AUTO to see how RISK handles those cases.

Begin with credit scoring and related practices. RISK can explain why many consumers are not wronged by being assigned a credit score, even if those scores are assigned via group membership or are the product of the credit agencies having blinded themselves to certain pertinent information. For many, the risk of being misclassified as a high-risk borrower is reasonably low. This is largely explained by the fact that credit agencies have ample incentive to score consumers accurately: the value of credit scores depends on their usefulness (i.e., predictive accuracy), and credit agencies are monitored by regulation (like ECOA) and consumers (who are empowered by regulation like ECOA and the Fair Credit Reporting Act). So, RISK has the resources to explain how the industry can blind itself to certain predictive data points (e.g., race, gender, ethnicity, and the like) while respecting our individuality: this can be done without exposing anyone to an unreasonable degree of risk.

Let's now turn to the case where the panicked parent looking for a doctor grabs the first person he sees wearing a white coat. In that case, the person in the white coat is exposed to some risk of being falsely identified as a doctor. However, the cost associated with misclassification is fairly low. Further, the context seems like it should push the threshold of acceptable risk up: that the parent is in an emergency situation makes less accurate, more effective methods for finding help more tolerable. (Compare: if the parent was in a scavenger hunt and used the same method to find a doctor, the person in the white coat might reasonably complain about the treatment he received from the parent.)

So, it seems that RISK is able to render the right verdict in Doctor, too. But, there may be a wrinkle. In Doctor, we are likely to have two intuitions: the person in the white coat was not wronged, and the person was not treated like an individual. Indeed, as Beeghly notes in the vignette, the father does not care whether he is grabbing this or that doctor. As it stands, RISK may stumble on this second intuition: it renders the judgment that the father does not fail to treat the doctor as an individual.

To iron this wrinkle out, we must adjust the language of our definition of RISK. I take it that the proper interpretation of the case is that while the father does not treat the person in the white coat as an individual, he does not fail to

pay sufficient respect their individuality, given the circumstances. Let us, then, change RISK as follows:[19]

> **RISK 2**: x fails to respect y's individuality if, and only if, x's treatment of y exposes y to a risk of being misclassified that y could reasonably reject.

We'd then have to update WRONG as follows:

> **WRONG 2**: it is wrong to fail to respect a person's individuality.

This version enables us to maintain the features that allowed RISK to deal properly with the machine bias case while getting both intuitive judgments about Doctor right.

Before we move on, it is worth asking whether GROUP (the account that Doctor posed the biggest problem for) could pull a similar move to deal with this case. Unfortunately for the account, it cannot. Suppose we revise GROUP as follows:

> **GROUP 2**: x fails to respect y's individuality if x's treatment of y is solely based on y's group membership.

This version of the idea still gets Doctor wrong because it violates what I take to be the core intuition of the case. That is, it (when accompanied with WRONG 2) renders the verdict that the father wrongs the person in the white coat. But this is clearly the wrong result.

4.2.2. New Problem Cases

One place where RISK 2 may run into some difficulties is in cases involving so-called "positive" stereotypes. Consider the following vignette, which Eidelson uses to motivate his account:

> **Sally.** Sally, who is of East Asian descent, auditions for her school orchestra. Sally plays the violin, but not seriously, and she is not particularly talented. Kevin, the orchestra director, thinks Sally performed poorly at her audition. But Kevin figures that Sally is probably a dedicated musician who just had a bad day, and selects her for the orchestra anyway. Kevin would not have made this assumption or selected Sally if not for her ethnicity and her sex.

In this case Kevin wrongs Sally by failing to respect her individuality. And yet, one might object, it is not at all obvious that he exposed her to a risk of being

19. For ease of comparison, the original RISK: x fails to treat y as an individual if, and only if, x's treatment of y exposes y to a risk of being misclassified that y could reasonably reject.

misclassified that she could reasonably reject. After all, if we suppose that Sally wants to make it into the orchestra, she might welcome Kevin's bias.

In what follows, I want to explore the options available to RISK 2 for responding to this objection. I think that there are roughly two options: *Option One* defends RISK 2 and says it can accommodate this case; *Option Two* concedes that this is a counterexample and modifies RISK 2 in response. I favor the latter option to the former but think that both are worth considering. Let's explore each in turn.

We begin our investigation into Option One by considering the grounds on which Sally might complain about Kevin's judgment. She could object that Kevin's propensity for falsely identifying her as good at the violin discounts her achievements; when she makes it into the orchestra this way, her achievement is worth less. Or, she might complain, Kevin's bias operates on the problematic assumption that Asians are a monolithic group, and that his accepting Sally into the orchestra serves to further entrench this problematic assumption. Or, perhaps, the idea that she's good at violin is bound up with negative ideas, such as the idea that she is unfit for leadership positions (Kim & Zhao 2014). We might use any of these grounds to make the case that Kevin has imposed too much of a cost on Sally in deploying a stereotype in his judgment of her.

I'm not very hopeful about Option One, but suppose for a moment that we can make a compelling initial case for it. This raises some interesting issues: What, then, about Prater? A bias towards positive outcomes for white people explains why he was identified as low-risk. Should we now worry that RISK 2 renders the judgment that he was wronged?

It seems, again, that we can go two ways on this. We could say that Prater was wronged or we could deny this. If we think he was not wronged, then we need to find a relevant difference between his case and Sally's. And perhaps we can. If, for example, Sally's complaint is that the bias that benefited her in the audition case is inextricably bound up with negative biases (like the idea that she is likely a poor candidate for leadership roles), then perhaps we can find this difference. If, on the other hand, we think Prater was wronged, we need to explain how this does not lead RISK 2 to give up one of its main benefits: its ability to explain how Jones and Borden were wronged to a greater degree than Prater.[20] This, too, can be accomplished. Were Prater wronged, it's still the case that Jones and Borden would have been at a higher risk of worse treatment, and so would arguably have been further above their threshold than Prater would have been.

Option One should perhaps be treated as a live option worthy of further exploration, but at present it does not seem especially promising. Trying to understand the wrong Sally suffered on the model of her being exposed to risk of misclassification involves quite a bit of shoehorning and doesn't feel like

20. Given what I have said previously, I would of course be reluctant to consider this option.

an especially comfortable fit. Let us, then, explore what I think of as a more promising, pluralistic approach to understanding the wrongs of biased judgments.

Suppose that Sally's case is a counterexample to RISK 2. How might we repair the account? Well, if it's a counterexample, it succeeds in refuting RISK 2's sufficiency claim (i.e., "x fails to respect y's individuality only if x's treatment of y exposes y to a risk of being misclassified that y could reasonably reject"). Suppose that, seeing this, we simply drop the claim and endorse what remains:

> **RISK 3**: if x's treatment of y exposes y to a risk of being misclassified that y could reasonably reject, then x fails to respect y's individuality.

Let's explore this option, beginning with the question of how accepting RISK 3 affects our analysis of Borden's, Jones's, and Prater's cases.

Accepting RISK 3 changes nothing about what we can say about the cases. Our analysis hinged on RISK's necessity claim. And, so, accepting RISK 3 maintains RISK's main attractive feature: its ability to capture the asymmetries between Borden's, Jones's, and Prater's cases.

But, then, how do we account for the Sally case? It seems that a modified version of AUTO can get it exactly right. To account for Sally's case, all we need is a version of AUTO's sufficiency claim (i.e., if x treats y as an individual, then the Character and Agency Conditions have been met), and this is because Kevin's failure to give reasonable weight to evidence of the ways Sally has exercised her autonomy in giving shape to her life, is how he fails to respect her individuality. Further, this claim is consistent with RISK 3. Accepting both claims just amounts to claiming that there is more than one way to fail to treat persons as individuals. On the other hand, AUTO's necessity claim (i.e., x treats y as an individual only if the Character and Agency Conditions have been met) has nothing to say about Sally, and it is inconsistent with RISK 3.[21]

If we embrace both RISK 3 and a modified version of AUTO we seem to have a powerful, though disjunctive, way of accounting for the wrongs engendered by wrongfully biased judgments. One might object to such an accounting on the grounds that it is not unified, to which I have two responses. First, it isn't obvious to me that a monolithic account based on RISK can't be made to work. (Though, as I have noted, I do not have much faith in it.) Second, a growing consensus in

21. To see this, suppose a case where the Agency and Character Conditions are met but the agent is exposed to an unacceptably high risk of misclassification. This is perhaps Borden and Jones's case. (How so? It might be the case that because of racist policing practices, Borden and Jones are—despite their short rap sheets—at high risk of being arrested in the future.) In such a case, judging them as high-risk is compatible with giving the evidence of the way they shaped their lives appropriate weight. But, it may also expose them to a high risk of misclassification. In such a case, AUTO's necessity claim judges that Borden and Jones are treated as individuals, whereas RISK 3 judges that they are not; thus, we must choose between the two claims.

the literature on stereotyping is that it was a mistake to think that there could be one monolithic explanation of the wrong of stereotyping based on the idea that wrongful stereotyping involves a failure to treat persons as individuals. Beeghly (2018) observes that a theory of wrongful stereotyping that avails itself of the idea that stereotyping involves a failure to treat persons as individuals owes us two things. One is that it must avoid the *problem of absurdity*, the problem of misdiagnosing as wrongful cases of stereotyping that are not wrongful (it must, for example, say that there is no wrong in Doctor). The other is that it must provide a conception of treating persons as individuals that can identify all cases of wrongful stereotyping as wrongful. Observing, as we have here, that monolithic accounts struggle to do both, she hypothesizes that it simply can't be done. As I think I have also shown, RISK isn't likely to recover everything on its own either. Insofar as we are attracted to the idea that stereotyping involves a failure to respect individuality, there is good reason to embrace a more pluralistic approach.

5. Closing Remarks

Let's take a step back and see what general features RISK 3 has such that it can get the right result in our case of machine bias and the old problem cases. RISK 3, unlike GROUP, INFO, and AUTO has a variable threshold of the degree of accuracy and tethers this threshold to factors that vary from context to context. The more general lesson is that some kind of an account with this feature is needed. RISK 3 is just one way of instantiating it.

I hope to have shown that giving an ethical accounting of machine bias is difficult because in indicting pernicious statistical judgments it is very easy to overgeneralize and indict statistical judgments that are not problematic. I also hope to have shown that this difficulty can be resolved if we accept an account—such as RISK, or some version of it—that tethers the degree of accuracy morally required of a statistical judgment to the stakes involved in making the judgment.

Acknowledgments

I am grateful to Rima Basu, Erin Beeghly, Molly Castro, Josh Mund, David O'Brien, Adam Pham, Alan Rubel, Ben Schwan, FIU's chapter of Phi Sigma Tau, participants at the 2019 Central Division meeting of the American Philosophical Association audience, and two anonymous reviewers. I am especially grateful to Jonathan Herington, my commentator at the Central APA, for extensive feedback.

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016, May 23). Machine Bias: There's Software Used across the Country to Predict Future Criminals and It's Biased against Blacks. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Antony, Louise (1993). Quine as Feminist: The Radical Import of Naturalized Epistemology. In Louise Antony and Charlotte Witt (Eds.), *A Mind of One's Own: Feminist Essays on Reason and Objectivity* (185–226). Westview Presss.

Barocas, Solon and Andrew D. Selbst (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.

Beeghly, Erin (2015). What is a Stereotype? What is Stereotyping? *Hypatia*, *30*(4), 675–691.

Beeghly, Erin (2018). Failing to Treat Persons as Individuals. *Ergo: An Open Access Journal of Philosophy*, *5*(26), 687–711.

Blum, Lawrence (2004). Stereotypes and Stereotyping: A Moral Analysis. *Philosophical Papers*, *33*(3), 251–289.

Corbett-Davies, Sam and Sharad Goel (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *CoRR*, *abs/1808.00023*. Retrieved from http://arxiv.org/abs/1808.00023

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (2017). Algorithmic Decision Making and the Cost of Fairness. *CoRR*, *abs/1701.08230*. Retrieved from http://arxiv.org/abs/1701.08230

Eidelson, Benjamin (2013). Treating People as Individuals. In Deborah Hellman and Sophia Moreau (Eds.), *Philosophical Foundations of Discrimination Law* (203–227). Oxford University Press.

Finlay, Steven (2014). *Predictive Analytics, Data Mining and Big Data*. Palgrave Macmillan.

Ghent, Andra C., Rubén Hernández-Murillo, and Michael T. Owyang (2014). Differences in Subprime Loan Pricing across Races and Neighborhoods. *Regional Science and Urban Economics*, *48*, 199–215.

Kim, ChangHwan and Yang Zhao (2014). Are Asian American Women Advantaged? Labor Market Performance of College Educated Female Workers. *Social Forces*, *93*(2), 623–652.

Lippert-Rasmussen, Kasper (2011). We Are All Different: Statistical Discrimination and the Right to Be Treated as an Individual. *The Journal of Ethics*, *15*(1-2), 47–59.

Northpointe Inc. (2016a). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity Performance of the COMPAS Risk Scales in Broward County*. Northpointe. Retrieved from https://www.semanticscholar.

org/paper/COMPAS-Risk-Scales-%3A-Demonstrating-Accuracy-Equity/cb6a2c110f9fe675799c6aefe1082bb6390fdf49

Northpointe Inc. (2016b). Risk Assessment Retrieved from https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

Obama, Barack (2013). Remarks by the President on Trayvon Martin Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2013/07/19/remarks-president-trayvon-martin

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen (2017). Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time. *Proceedings of the National Academy of Sciences*, *114*(41), 10870–10875.

Rice, Lisa and Deidre Swesnik (2013). Discriminatory Effects of Credit Scoring on Communities of Color. *Suffolk University Law Review*, *46*, 935–966.

Turner, Margery Austin, Rob Santos, Diane K. Levy, Doug Wissoker, Claudia Aranda, Rob Pitingolo, and The Urban Institute (2012). Housing Discrimination against Racial and Ethnic Minorities. Retrieved from https://www.huduser.gov/portal/Publications/pdf/HUD-514_HDS2012.pdf

United States Executive Office of the President (2014). Big Data : Seizing Opportunities, Preserving Values. White House, Executive Office of the President.

State v. Loomis, 881 N.W.2d 749 (Wis. 2016).