A PHILOSOPHICAL ANALYSIS OF AI AND RACISM

LEL JONES

ABSTRACT

This paper addresses the problem of racism against Latinx and Black people in Artificial Intelligence (AI) and offers possible solutions. This ethical analysis is necessary because with a dramatic increase in the production of AI, the way we use it is critical in eliminating its current perpetuation of racism. I offer evidence of the current perpetuation of racism through AI by analyzing its use in banking and in law enforcement. I argue that the current way we produce and use AI needs to be seriously reconsidered and reinforce this argument with the use of Rawls' theories of justice.

I. INTRODUCTION

The rise in Artificial Intelligence (AI) has birthed fears of an *Age of Ultron*-esque doomsday. While the potential of world-wide human destruction is scary, it is not an immediate problem. In this paper, I argue that a more urgent issue concerning AI is the way it perpetuates racism. I will examine the causes of this issue and argue why we must address it through ethical analysis. The rapid development of AI poses exciting opportunities for productivity, as AI can often accomplish assigned tasks quickly and inexpensively. Within an economic framework of capitalism, this means that AI can outperform humans. Unfortunately, AI's extreme efficiency also contributes to the continued marginalization of people of color, especially Latinx and Black people (LBP).

Instances of racial discrimination in U.S. business and agency AI systems are already demonstrable. Google's facial recognition software "recognized" LBP as gorillas due to the lack of LBP in their original facial survey.[1] AI is absorbing jobs from employees in fields with disproportionately many LBP and is being redistributed to the tech industry where only 7.5% of employees are LBP.[2] Throughout the paper we will examine how these examples expose an issue of urgent ethical concern in part by employing John Rawls's "veil of ignorance." In Section II, we will address the conceptual issue of how it is possible for AI to be racist. After that, we will consider some examples of discrimination. In Section III, we will examine banking AI charging LBP rates almost double that of white people with the same FICO score.[3] In Section IV, we will examine how the U.S. police use the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS).[4] This program has been proven to misinterpret black people as high risk reoffenders twice as often as white people and misinterpret white people as low risk reoffenders twice as often as black people.[5] In Section V, we will consider the objection that these are necessary consequences of

1 Jessica Guynn, "Google Photos Labeled Black People 'Gorillas,'" *USA Today*, July 1, 2015, http://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/.

2 "Diversity in High Tech," U.S. Equal Employment Opportunity Commission, accessed January 29, 2020, http://www.eeoc.gov/eeoc/statistics/reports/hightech/.

3 Robert P. Bartlett et al., "Consumer Lending Discrimination in the FinTech Era," *SSRN Electronic Journal* (2017), 10.2139/ssrn.3063448.

4 Julia Angwin et al., "Machine Bias," *ProPublica*, March 9, 2019, http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

5 ProPublica analyzed "risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm."

AI advancement, and human progress should not be stopped because a minority of people are harmed. We will respond to this objection in Section VI by demonstrating that this is an appeal to a weaker kind of consequentialism that considers too few factors and, as critiqued by Rawls, does not capture the relevant elements of justice.

II. HOW CAN A MACHINE BE RACIST?

It is important to briefly address the common reaction to the problems identified above: that the AI itself is racist; we have no way of knowing what is going on inside the "black-box." This view often results in objections to any implementation of AI because it could be prejudice. However, this "black-box" argument could also be applied to people. We can hook someone up to a machine to see that their neurons are firing correctly; we can do a similar test on AI by examining its coding. It is not the numbers inside an AI that are racist, rather it is the methods that people use when coding them—particularly through the survey samples used—and the institutions within which they operate. This is an important distinction to make because without it, we are able to shift the blame from ourselves onto AI. This objection could have serious consequences, as AI has already shown that it has important benefits. When we accept that these issues are a result of human error, we can begin to take the necessary steps to resolve the problems—as opposed to giving up on AI all together.

III. RACISM IN BANKING AI

According to a University of California Berkeley (UCB) study titled "Consumer-Lending Discrimination in the Era of FinTech," banking AI charges Latinx and Black people an average of 7.5 basis points higher interest rates than it charges white people.[6] Financial technology (fintech) in banking was initially predicted to eliminate, or at least reduce, discrimination in U.S. banking and loaning. When studies like UCB's concluded that fintech programs like Quicken and SoFi discriminate to a similar degree as their human counterparts, many claimed that it was the AI doing the discriminating. As previously stated, AI itself is not prejudiced. Rather, AI is infiltrated with the racist beliefs of the people and institutions that contribute to its development and programing. Deferring the blame for racist outcomes is not limited to misplacing it on AI. Programmers and institutions point to other factors as being responsible for racist outcomes. In the case of fintech disproportionately surcharging LBP—despite their FICO scores being the same as other applicants—companies responded by citing other demographic data they used as the criterion for disparate interest rates.

6 Bartlett et al., "Consumer Lending Discrimination."

This data included education, the urgency that the money was needed, and more. Factors like education re-marginalize LBP, because they are institutionally disenfranchised through white Eurocentric curricula, they are minimally represented in many fields of study, and they are not accepted into colleges and universities at proportional rates to their white counterparts, among other issues.

One counterargument to this claim is that the bank is just trying to make money, not promote social justice, so it cannot be held accountable for racist issues within the U.S. This is where the introduction of AI has a meaningful impact. AI is goal oriented and performs tasks in a hyper-efficient way; that is why we build them. Since bankers are underwriting AI with the same goal they have without AI—which is to give the most expensive loans possible—AI will find ways to do that more efficiently than bankers. Although AI is at a similar quantitative discrimination level as the bankers now, it will continue through the path of least resistance to yield the most expensive rates. AI is likely to find that path through LBP. The software is able to explore avenues of data—more comprehensively than humans can—that statistically make a person more likely to default. This can quickly spin out of control. AI can survey a quantity of factors like education, and for a demographic that is institutionally disadvantaged the result will likely be a continual increase in their rates. Even if one denies 7.5 points of discrimination as a weighty-enough rights violation to warrant (more) limits on banks, AI is likely to significantly increase this number.

This counterargument can also be addressed from a Rawlsian perspective. In his book, *A Theory of Justice*, Rawls argues that justice must not be separated from the workings of the economy.[7] Rawls famously argues that we should look at society from the "original position," where we are behind a "veil of ignorance" that blinds us from knowing exactly who, where, and when we are existing in society. We may be rich or poor, educated or uneducated, religious or secular, etc., but we do not know when we decide the basic rules of society. Rawls argues that anyone in the original position would—out of purely rational self-interest—set the basic rules of society to maximize the benefits to those who are worst off, as they could end up being one of them.

Applied to the current case, Rawls would have a clear objection to the idea that banks should be solely concerned with making money, regardless of social justice concerns. The fact that our use of AI is harming a group of already marginalized people is unjustifiable. According to Rawls, there should be limits that "follow from the

7 John Rawls, *A Theory of Justice* (Cambridge: Belknap Press of Harvard University Press, 1971), 230.

priority of justice over efficiency and the priority of liberty over social and economic advantages."[8] In the case of banking AI, we are seeing the opposite of this. Efficiency is being prioritized over justice, with a significant cost to LBP. The fact that our current use of AI in banking is hurting LBP creates further friction with Rawls' theory.

This Rawlsian response becomes even more powerful when one considers that AI is likely to begin finding that other institutionally disadvantaged identities—such as being disabled, LGBTQ, etc.—play a role in likeliness to default. There are many people who are, unjustly, more vulnerable than others. When AI is programmed to value efficiency over ethical consequences, it takes advantage of marginalized populations quicker and to a higher degree than humans can. So, although one may be sensitive to allowing human bankers to run their business in the most profitable way, the increased consequences of AI running businesses for profit necessitates limitations. Outside of an economic framework, the consequences of misusing AI extend into other institutions, in this case law enforcement. By looking at examples of this kind of injustice, we avoid economic objections all together.

IV. RACISM IN LAW ENFORCEMENT AI

The U.S. police force has implemented the use of COMPAS to evaluate the likeliness of an arrested person to reoffend, based on an analyzation of data from a survey they were instructed to take. This technology was said to be implemented to reduce bias in law enforcement, yet it continues to reinforce negative stereotypes about LBP being recurrent criminal offenders. Similar to the banking example, it would be a mistake to say the computer program is prejudiced itself. Instead, we must look at the factors that create a biased program. With this example, we will examine the dangers of the assumption that the technology we create will be better at circumventing discrimination than humans. Under this mentality, it is possible that we will fail to properly evaluate AI outcomes. Analyzing the effects of these programs requires special access that few have, so our perception of people's ability to create AI that mediates complicated human affairs should remain critical.

The impacts of COMPAS are of great consequence because judges use the COMPAS scores to assist in determining sentences. A study from ProPublica shows that COMPAS misinterprets black people as high-risk reoffenders twice as often as it does white people, and misinterprets white people as low-risk reoffenders twice as often as it does black people.[9] When judges consider COMPAS, they will potentially create harsher sentences for low-risk black people and

8 Rawls, *A Theory of Justice*, 230.
9 Angwin et al., "Machine Bias."

lighter sentences for high-risk white people. After ProPublica's study was published, Dartmouth University conducted a study revealing that COMPAS was no better at judging recidivism than random volunteers from the internet. After these studies were published, many courts suggested using caution with COMPAS scores. This would not have happened if Dartmouth did not conduct research about the police program. This research was unprompted by the department of law enforcement responsible for creating COMPAS. That department simply assumed it would work properly, and the result was an unknown quantity of injustices to black people. This example highlights how urgently we need to resolve such problems. In addition to relieving direct injustice of undeserving sentence length and intensity, the mistakes made by COMPAS will skew our data. Statistics will start to show that more black people are more likely to reoffend than white people, and those statistics will be used in COMPAS calculations. This would result in an acceleration of its original mistakes and the production of statistics from false information. A person profiled by COMPAS will find their interactions with the criminal justice system systematically weighted against them.

Again, a Rawlsian perspective is instructive here. Rawls argues that behind the veil of ignorance, everyone—regardless of their conception of the good life—would want access to what he calls "primary goods." These primary goods are the basic constituents of a satisfying life in any possible social structure. First and foremost among these primary goods is the sense that one's life has value, and their projects and convictions are worth carrying out. As Rawls succinctly puts it, "perhaps the most important primary good is that of self-respect."[10] Because self-respect (or self-esteem) is central to any possible conception of the good life, Rawls claims that "parties in the original position would wish to avoid at almost any cost the social conditions that undermine self-respect."[11] COMPAS and programs like it pose a significant threat to the self-esteem of marginalized populations. An abundance of research concludes that experiencing repeated instances of discrimination significantly lowers one's self-esteem. For example, Subadra Panchanadeswaran and Beverly Araujo Dawson, have found that discrimination and stress causes lower self-esteem in Dominican women.[12] Ethan H. Mereish has done research that suggests there is a significant association between discrimination and self-esteem

10 Rawls, *A Theory of Justice*, 386.

11 Rawls, *A Theory of Justice*, 386.

12 Subadra Panchanadeswaran and Beverly Araujo Dawson, "How Discrimination and Stress Affects Self-Esteem Among Dominican Immigrant Women: An Exploratory Study," *Social Work in Public Health* 26, no. 1 (2011): 60–77, 10.1080/10911350903341069.

in African American men.[13] These results are replicated even when analyzing different kinds of discrimination. For example, Vickie M. Mays and Susan D. Cochran found similar results when looking at the LGBTQ community.[14] Given that these empirical studies demonstrate how discrimination undermines self-esteem, anyone sitting behind Rawls's veil of ignorance would not tolerate any program—like COMPAS—that routinely undercuts this primary good. When creating AI, we must plan for rigorous testing and not restrict LBP's access to the important primary good of self-respect.

V. COUNTERARGUMENT BY APPEAL TO PROGRESS

The idea that we must reevaluate how AI is implemented receives pushback primarily from an appeal to progress. This argument—that we should continue developing AI without properly studying the effects of its programing—is essentially a consequentialist argument. It states that human progress is worth the administration of some harm to a minority of people. The fruits of our advancement in AI will outweigh the costs it will incur. Arguments like this often cite previous technological revolutions—for example, the computer. While having caused some harm, like eliminating a significant amount of jobs, the computer has by far been worth it. It has created thousands of jobs, revolutionized communication, and reduced the gap between people and information. We would likely not be where we are today if we stopped to remedy every injustice enacted in the name of computer technology.

VI. CONSEQUENTIALISM WITHOUT JUSTICE

While consequentialism has merits, the kind of consequentialism used in this argument has been criticized by philosophers like John Rawls, claiming that it struggles to capture the nuances of injustice.[15] Limiting our analysis to the quantitative consequences of AI oversimplifies the repercussions. The argument merely considers the number of people who experience good or bad consequences from

13 Ethan H. Mereish et al., "Discrimination and Depressive Symptoms Among Black American Men: Moderated-Mediation Effects of Ethnicity and Self-Esteem," *Behavioral Medicine* 42, no. 3 (2016): 190–96, 10.1080/08964289.2016.1150804.

14 Vickie M. Mays and Susan D. Cochran, "Mental Health Correlates of Perceived Discrimination Among Lesbian, Gay, and Bisexual Adults in the United States," *American Journal of Public Health* 91, no. 11 (2001): 1869–76, 10.2105/ajph.91.11.1869.

15 George Sher, *Ethics: Essential Readings in Moral Theory* (New York: Routledge, 2012), 263.

AI, as opposed to how good or bad the effects are.[16] More efficient transportation resulting from AI is a good that could easily benefit the majority of people, but if it comes at the cost of unequal treatment and harm to a minority, it is not ethically warranted. When we only address the number of people who experience consequences, we ignore problems like the perpetuation of racism through AI because it targets those in the minority. For example, while bank owners will benefit to some degree by obtaining higher profits, LBP will suffer to a higher degree from exploitation. In this way, this consequentialist framework does not make room for the prioritization of justice. This is especially problematic because—as my examples show—LBP injustices resulting from AI will continue to increase if ignored.

An additional response to this argument is that it assumes progress to be only technological or economical. It reduces the relevant consequences to things such as profits and excludes the value of equal treatment. Progress in racial justice should be considered equally valuable, if not more valuable. Furthermore, this argument tends to look at those who are better off when assessing progress: banks, tech companies that earn money from AI, and those who can afford AI. If we are progressing towards an ideal society, it seems unwise to use the most privileged people as a metric for progress. Instead, we should focus on those facing obstacles that push them further away from the ideal, as it gives us a more accurate representation of the distance between the present and the objective. For example, Rawls suggests that we build our societies behind a veil of ignorance where we do not know who, when, or where in society we will be.[17] We do this because when this information is withheld, we will make decisions that benefit the worst off in society, because we could be among those who are the worst off. In this context, the worst off are often those who are subject to injustice, and—when it comes to AI—that is most often LBP. Behind the veil of ignorance, we would make the decisions necessary to eliminate racism in AI because behind the veil, we are aware that we could be harmed by it.

An objection to the prioritization of justice in Rawls' framework is that if we always halted progress to correct every injustice, we would never move forward. AI can efficiently perform certain tasks that will ultimately save lives, including robotic surgeries with higher success rates and advanced AI cars that limit accidents. This argument, however, relies on the false dichotomy between AI as we currently use it and no AI advancement at all. We can still move forward with AI development

16 Of course, there are forms of consequentialism that could possibly produce a verdict in favor of justice like Mill's qualitative Utilitarianism. John Stuart Mill, *Utilitarianism* (Indianapolis: Hackett, 2002).

17 Sher, *Ethics*, 387.

so long as we take provisions to minimize injustice. This would look like a tech industry with proportional representation of LBP, software data samples that include an appropriate amount of data from all races and ethnicities, rigorously testing AI before it is implemented, and limiting the disparity between the treatment and charging of different identities in economic settings. The same person objecting to the bank example would interject here by saying that it is unjust to put limitations on the financial margins procured by AI because it violates the company's right to profit. However, the company only maintains this right when doing so does not infringe on the rights of others. I would argue they infringe on others' rights by exploiting the disadvantaged status of LBP. Even if one disagrees with this by saying the margins of discrimination are not large enough to qualify as a rights violation, there is no inconsistency in acknowledging that. Because AI profits are likely to intensify injustices, we should place limits on those profits and methods. We could take the strong stance that we have to limit the magnitude of differential treatment and charges between white people and LBP in all circumstances. Or, we could take the weaker stance that because injustice increases in the economic framework of AI, we will have to limit the exploitation of LBP for profit. While I maintain the strong stance, the weak stance is still compatible with the position that we must resolve the way we build and use AI to cease the perpetuation of racism.

VII. CONCLUDING REMARKS

The addition of AI in business does not guarantee the elimination of bias and discrimination. On the contrary, there are many examples of how our current use of AI is perpetuating racism in financial and criminal institutions. While we may sacrifice some speed in technological progress, we must intensify our scrutiny when analyzing how AI will impact people of color, especially LBP. The argument that we should not sacrifice speed in technological progress because it is worth a minority of harm is a weak consequentialist argument that fails to capture elements of justice, multiple kinds of progress, and how harm is distributed.

The proposed steps to eliminate racist issues in our use of AI will not halt the use of AI altogether. Ideas for further thought include how AI could benefit LBP when it is executed properly. AI could be used to administer prejudice tests to reduce implicit bias, it could map environmental hazards for LBP in higher risk areas, and it could even improve access to transportation for LBP. Addressing and resolving the racism we currently face with AI does not mean that AI progress will or should come to a stop.

Lel Jones will be receiving their B.A. in Philosophy with a concentration in Honors and Ethics, Politics, and Law from California State University, Sacramento this Spring 2020. Their philosophical interests are primarily in bioethics, neuroethics, feminist philosophy, and epistemology.