RACIST ROBOTS

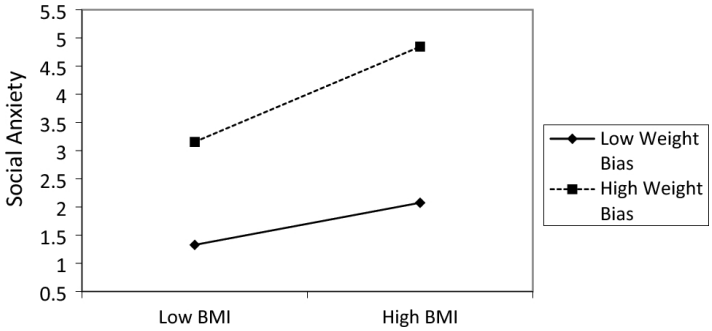
HOW ALGORITHMS' OBJECTIVE DECISIONS ECHO SUBJECTIVE PROBLEMS

RACIST ROBOTS: HOW ALGORITHMS' OBJECTIVE DECISIONS ECHO SUBJECTIVE PROBLEMS

Ethan Nguyen

It was a warm spring day on the way home several years ago that I found myself talking to my father about selective algorithms, tools used by computers to make decisions without human input. He was a seasoned and professional computer scientist who'd worked as a software engineer longer than I'd even been alive. He believed that these algorithms were a part of the development of humans as a species, presenting ways to efficiently make choices free from human limitations of emotion. I told him an amusing story about an algorithm trained for interpreting sentences. Over time, it had become sexist, assigning genders to subjects without any existing in the first place: the pool of sentences it was trained off included sentences with sexist tendencies. His expression suddenly hardened as he explained to me that this was not possible: machines are unemotional things, and anyone who claims that they can be prejudiced is reading too far into something indifferent to the petty differences humans struggle with.

In recent years, algorithms, or sets of instructions for a computer to achieve a goal, have increasingly become relevant. They have been used to filter people for employment, decide who should receive medical assistance, and more. In theory, computer algorithms are an ideal solution to a problem humans have long struggled to overcome: bias. While humans have always struggled to overcome this, a recent development in our understanding of it has been implicit bias. Implicit bias is a form of bias that every human struggles with that causes them to automatically and unconsciously judge others in prejudiced ways, regardless of their actual social beliefs. For example, a study analyzing the relationship between social anxiety and weight bias found a strong correlation between social anxiety and intrinsic weight bias in individuals with high BMIs, despite finding low correlation between implicit and explicit weight biases, demonstrating that people were being influenced by biases they did not consciously or knowingly acknowledge themselves as having (Kaplan et al., 2023).



The relationship between weight bias, social anxiety, and BMIs. People with high BMIs and intrinsic weight bias were found to have higher rates of social anxiety. These people were not associated with explicit weight bias, showing how internalized beliefs can be as harmful as explicit biases (Kaplan et al., 2023).

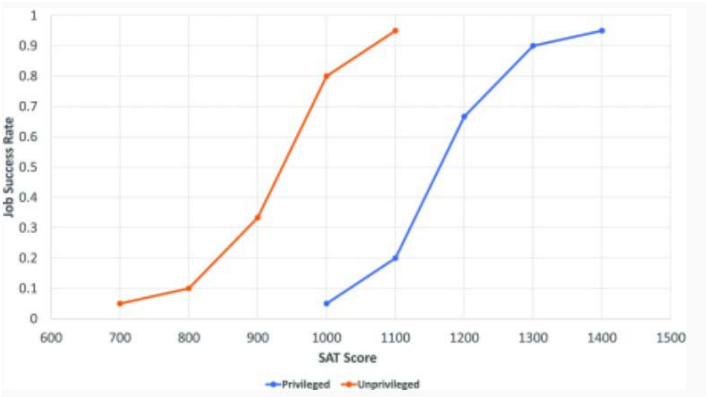
Intrinsic bias has caused many to worry about mitigating bias when it seems that humans are inherently biased. How can we prevent employers from inadvertently factoring gender into their decisions about how qualified a candidate is? How do we prevent educational institutions from factoring race into decisions about students? This is where computers come in. Computers have the power to make analytic decisions without human input on scales previously unknown, allowing less resources to be allocated to making more decisions. Using data analyzing various aspects of dataset, algorithms can be created. They can compare a person's aspects against data from previous people used in training to make objective predictive decisions based on patterns that the algorithm recognizes. Because computers are not vulnerable to emotional biases the way humans are, they should be able to make these decisions without considering irrelevant aspects of a subject, leading them to objective decisions that are as factually-influenced as possible.

In judicial systems, risk assessment algorithms (RAIs) have been heralded as the solution to human bias. They are used to decide whether or not someone detained for a crime should be jailed

"THEY HAVE BEEN USED TO FILTER PEOPLE FOR EMPLOYMENT, DECIDE WHO SHOULD RECEIVE MEDICAL ASSISTANCE, AND MORE."

before trial, a decision made on the basis of whether or not someone is evaluated to be at high or low risk of committing another crime while awaiting trial (Angwin et al., 2016). They've become a piece of the United State's ongoing prison puzzle: the country, despite consisting of only 4% of the global population, contributes 20% to the world's incarcerated population (Barabas et al., 2019). Many have had hopes that these algorithms would help solve the US's overflowing prison problem by efficiently and objectively choosing high-risk individuals while letting low-risk individuals stay free, minimizing both how many people are in jail and how many go free and commit crimes (Barabas et al, 2019). Because these algorithms are objective, they would also ensure that evaluation of whether or not someone was high or low risk would be as free from prejudice as possible.

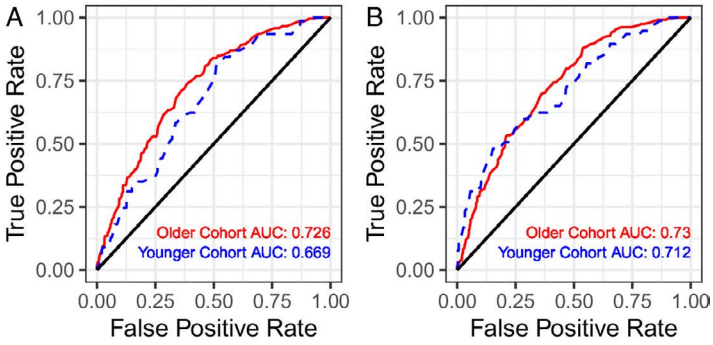
Unfortunately, predictive algorithms have proven themselves to be far from perfect. Oftentimes, bias can arise in algorithms because they are trained off datasets that were influenced by human biases, and the algorithms begin to reflect this (Pessach, 2023). For example, a hypothetical algorithm is trained to prioritize patient treatment based on a variety of factors, including the amount of money that was spent to treat similar patients, because more money spent on treatment correlates with more serious afflictions. However, in this hospital, male patients are more likely to have their symptoms taken seriously and have money spent on their treatment. An algorithm trained from this data correlates male patients with more treatment, and evaluates male patients to be higher priority for treatment than non-male patients. It has also been found that even without biased training sets algorithms are vulnerable to bias anyway, such as from incomplete data causing misrepresentation of populations, leading to algorithms making decisions about populations that don't actually reflect the populations properly (Pessach, 2023). Algorithms can even draw illegitimate conclusions about legitimate data, and several studies have shown that it is not possible to resolve algorithmic bias without compromising accuracy (Pessach, 2023). Dana Pessach, a researcher at Tel Aviv University specializing in data science and machine learning, presents a hypothetical scenario where an algorithm is used for hiring that checks candidates' SAT scores. It finds that, generally, lower SAT scores lead to worse job performance, and so favors candidates with higher SAT scores. In the social context, however, she splits people into privileged and underprivileged groups based on their upbringings. In this scenario, people from underprivileged backgrounds with less educational resources score lower scores on the SAT than their counterparts. However, when comparing two candidates from differing backgrounds with the same rating on an arbitrary "job success" scale, the underprivileged candidate would generally have the lower SAT score. The algorithm, however, does not make a distinction between the two groups, and favors privileged candidates over underprivileged ones (Pessach, 2023).



Pessach's model of privileged candidates versus unprivileged candidates. Overall, lower SAT scores correspond with worse job performance, but candidates in the unprivileged category had worse SAT scores for the same job success rates (Pessach, 2023).

Predictive algorithms even lose their power over time. For example, a group of researchers from Carnegie Mellon, the University of Pennsylvania, and Harvard conducted research on cohort bias, a type of bias that arises from categorizing groups of people. The researchers found that, as a particular algorithm's training dataset aged, the algorithm's predictions would become less accurate and increasingly biased. This happened because using old data caused age groups—or "cohorts"—

to form between the people used to train data and the people that the algorithm was making predictions for. The more separated the training group's ages were from a particular subject, the worse the algorithm's ability to predict outcomes was. In other words, as they aged, datasets became worse at predicting outcomes for people in the current social context, amplifying biases (Montana et al., 2023).



True positive rates for two sets, analyzing an algorithm's accuracy based on age cohorts. Older cohorts (those closer in age to the ages of people the algorithm's dataset was trained off of) had higher true positive rates than younger cohorts (Montana et al., 2023).

"SEVERAL STUDIES HAVE SHOWN THAT IT IS NOT POSSIBLE TO RESOLVE ALGORITHMIC BIAS WITHOUT COMPROMISING ACCURACY. "

The high expectations for these seemingly objective algorithms has led to a problem with their implementation: these algorithms are not closely monitored for their actual robustness, under the assumption that they simply work. This issue has been reflected in academia, where the assumption has led to unrobust research that was originally meant to analyze the effectiveness of RAIs. In a 2013 study analyzing 47 other studies reviewing various risk assessment tools, researchers found that most misinterpreted their data in favor of the algorithm, and few provided justification for the usage of some statistics, sometimes simply claiming that their results were agreeable with previous studies. The researchers analyzed each studies' interpretation of "area under the curve" (AUC), a statistic that would indicate the probability that someone who committed a crime receives a higher risk rating than someone who does not commit a crime. Only 16 did a statistical analysis of the area under the curve, and only two of these correctly interpreted the meaning of the AUC. The remaining 14 most commonly misinterpreted AUC as the proportion of individuals whose outcomes were correctly predicted (Singh et al., 2013). In practice, the assumption has led to RAIs' decisions being treated more as fact than evidence. Often, defendants cannot challenge the risk score they are assigned. Hearings do not investigate the underlying factors used to determine a score, and information on what factors the algorithm used to determine a score is often inaccessible (Angwin et al., 2016). Because of the misconception that these tools make objective decisions, little thought is put into allowing their results to be challenged. However, we cannot afford to let these algorithms operate unquestioned: the consequences can be as serious as altering the course of people's lives.

An independent investigation conducted by ProPublica found that an RAI named COMPAS, which is popular nation-wide, was twice as likely to label black defendants who did not re-offend as high-risk, yet twice as likely to label white defendants who did as low risk (Angwin, 2016). This means that many black people are finding themselves with more serious charges than white people because they are more likely to be labeled as high risk for more minor crimes. In 2014, Sade Jones, a then 18-year-old, attempted to steal a children's bike left on the lawn of a suburban home. Despite the fact she had no criminal history, she was rated as medium risk by COMPAS. Her bond was raised from a recommended $0 to $1,000, and she was given a felony burglary charge (Angwin, 2016). Her story is only one of many who have found themselves with inflated risk ratings and charges based on their skin color. A few months before Jones' arrest, a white man named Vernon Prater was arrested for shoplifting. Despite several previous armed robbery charges, he was rated low risk. Later, he broke into a warehouse and stole thousands in goods (Angwin et al., 2016).

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

An investigation of the predicted outcomes of people rated by COMPAS versus their actual outcomes. ProPublica found that black people were almost twice as likely as white people to be labeled as high-risk but not re-offend (Angwin et al., 2016).

"THE CONSEQUENCES OF A WRONG DECISION CAN DESTROY PEOPLE'S LIVES."

Clarence Okoh, a political scientist with former ties to the NAACP and the New York Civil Rights Bureau, describes "black coding," a concept he defines as the usage of algorithms or other means to sabotage civil rights protections to protect existing prejudiced systems. "Left unchecked, the cumulative effect of black coding can result in a collection of practices that effectively return racially marginalized communities to second-class citizenship," said Okoh, in an opinion piece he first published in 2022. He argues that these algorithms are the modern day iteration of a long series of practices used to prevent black Americans from gaining equal social status to their white counterparts, with the original black codes starting soon after the American Civil War (Okoh, 2022). This may not necessarily be true: risk assessment algorithms were originally meant to forward society by providing an unbiased means to find high-risk criminal offenders, but still, they instead reinforce the prejudiced systems they were meant to eliminate. Technology has advanced rapidly in the past ten years: more memory is crammed into smaller containers, games scenes go from low-resolution images to meticulously calculating each ray of light, and generative AI has evolved from existing at the butt of jokes to being the subject of hot ethical debate. Yet, risk assessment algorithms drag their feet and fall behind sluggishly: Okoh's opinion piece calls out their shortcoming and misuse today, echoing the 2013 study nearly a decade later.

Some researchers that recognize that biases exist in risk assessment algorithms do not advocate for doing away with these algorithms. The researchers behind the paper on cohort bias agreed that cohort bias presented a detriment to the merit of these algorithms. Despite this, the researchers did not advocate not to use these algorithms, arguing that judges, who would otherwise make risk assessment decisions, are just as prone to cohort bias as the algorithms were. They instead advocated for reform. On a social level, they believed that investigation needed to be done to see how cohorts arise and attempt to account for these social forces in writing algorithms, alongside accounting for the age of datapoints in the algorithms themselves. The researchers did not comment on other forms of bias, but their methods of reform can help counter other types of bias as well. For example, the researchers believed data from algorithms should be updated regularly to ensure accurate and better representations of a population: they found that many risk assessment algorithms could go over a decade without updating their datasets. Investigating how social forces influence sources of biased data—such as why risk assessment algorithms are so biased against black people in the first place—could also shed light on how we can minimize these issues in the first place.

In the meantime, if we are to continue to use these algorithms, we have a duty to recognize they are flawed. The misconception that these algorithms are objective and unbiased is dangerous, because it means that we do not question the undeniably questionable decisions that these algorithms make. The consequences of a wrong decision can destroy people's lives, and to do so when we could easily have done differently is a crime. People are complicated and hard to judge: algorithms or not, it will never be possible to determine what they might do with 100% certainty. Mistakes will happen, and so will injustices as we continue to struggle with how we make just decisions in court. However, what we can do is minimize this by remaining critical, questioning the decisions made by predictive algorithms, and slowly creeping to a better future.

REFERENCES

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barbaras, C., Dinakar, K., Doyle, C. (2019, July 17). The Problems With Risk Assessment Tools. *The New York Times.* https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html

Kaplan, S. C., Butler, R. M., & Heimberg, R. G. (2023). The Relationship Between Body Mass Index, Implicit Weight Bias, and Social Anxiety in Undergraduate Women. *Cognitive Therapy & Research, 47*(5), 761–771. https://doi-org.colorado.idm.oclc.org/10.1007/s10608-023-10404-6

Montana, E., Nagin, D. S., Neil, R., & Sampson, R. J. (2023). Cohort bias in predictive risk assessments of future criminal justice system involvement. *Proceedings of the National Academy of Sciences of the United States of America, 120*(23), 1–9. https://doi-org.colorado.idm.oclc.org/10.1073/pnas.2301990120

Okoh, C. (2022, March 8). The Dilemma of Black Coding: Assessing Algorithmic Discrimination Legislation in the United States. *Transatlantic Policy Quarterly.* http://turkishpolicy.com/article/1112/the-dilemma-of-black-coding-assessing-algorithmic-discrimination-legislation-in-the-united-states

Pessach, D., Shmueli, E. (2023). A Review on Fairness in Machine Learning. *ACM Computing Surveys, 55*(3), 1–44. https://doi-org.colorado.idm.oclc.org/10.1145/3494672

Singh, J., Desmarais, S., Van Dorn, R. A. (2013). Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review. *Behavioral Sciences & the Law, 31*(1), 55–73. https://onlinelibrary.wiley.com/doi/10.1002/bsl.2053