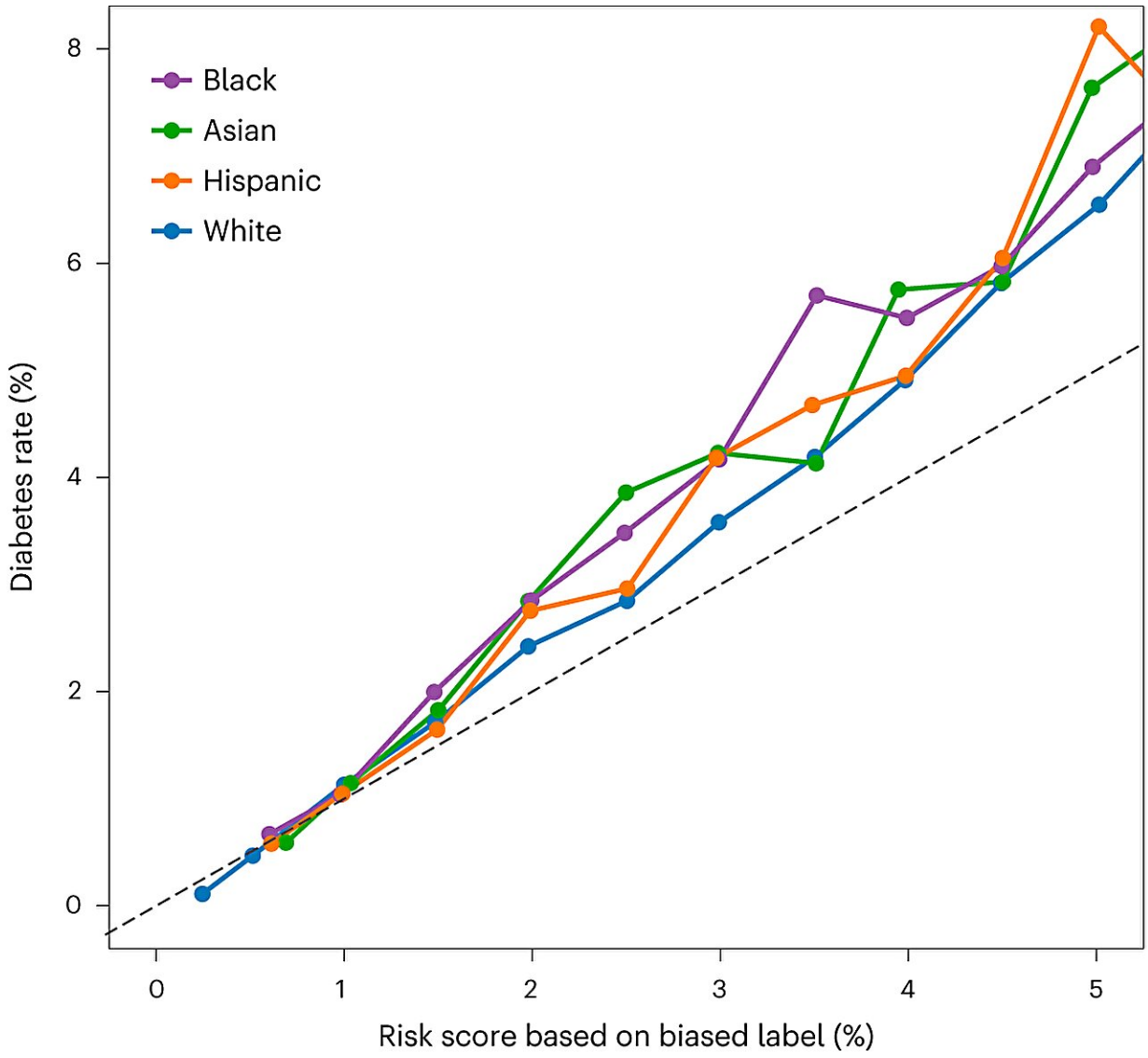
Rethinking algorithmic decision-making based on 'fairness'

July 31 2023, by Monica Schreiber



The impact of label bias on calibration. Estimated diabetes risk against observed diabetes rates across demographic groups. The diagonal dashed lines represent

hypothetical risk scores that are perfectly calibrated to empirical diabetes rates. In contrast to the risk scores presented in Fig. 1, which predict the prevalence of diabetes from blood tests and doctors' diabetes diagnoses, here the models are only trained to predict a doctor's diabetes diagnosis. The model inputs are age, race/ethnicity and BMI. Probably due in part to racial and ethnic disparities in health care access, predicting a doctor's diabetes diagnosis introduces bias into the model when compared against the results from the combined label. We observe that Asian, Black and Hispanic patients have higher true diabetes risk than white patients with the same nominal risk under the model. Credit: *Nature Computational Science* (2023). DOI: 10.1038/s43588-023-00485-4

Algorithms underpin large and small decisions on a massive scale every day: who gets screened for diseases like diabetes, who receives a kidney transplant, how police resources are allocated, who sees ads for housing or employment, how recidivism rates are calculated, and so on. Under the right circumstances, algorithms—procedures used for solving a problem or performing a computation—can improve the efficiency and equity of human decision-making.

However the very standards that have been designed to make algorithmic decisions "fair" might actually be entrenching and exacerbating disparities, particularly along racial, ethnic, and gender lines. That's the thrust of "Designing Equitable Algorithms," a paper published this week in *Nature Computational Science* by Stanford Law Associate Professor Julian Nyarko, Executive Director of the Stanford Computational Policy Lab Alex Chohlas-Wood, and co-authors from Harvard University.

With the proliferation of algorithmically guided decision-making in virtually all aspects of life, there is an increasing need to ensure that the use of algorithms in making important decisions is not leading to unintended negative consequences, according to the authors.

"A decision-maker can define criteria for what they think is a fair process and strictly adhere to those criteria, but in many contexts, it turns out that this means that they end up making decisions that are harmful to marginalized groups," said Nyarko, who focuses much of his scholarship on how computational methods can be used to study questions of legal and social scientific importance.

Nyarko cited diabetes screening as an example. "Algorithms are used as a first filter to determine who receives further tests. We know that, given a certain BMI and age, patients who identify as Asian tend to have higher diabetes rates than those who do not identify as Asian. An algorithm that has access to a patient's race can use this information and be appropriately more lenient in its referral decision if the patient identifies as Asian.

"However, if we insist on race-blind decision making, we make it hard for the algorithm to use that information and adjust its predictions for Asian patients. Ultimately, this means that the race-blind algorithm, although it might be 'fair' in a technical sense, excludes from further testing some Asian patients with a demonstrably high diabetes risk. A similar trade-off between what we might call a fair process and equitable outcomes applies to most popular fairness criteria that are frequently used in practice."

Results like these are well-known in the literature on algorithmic decision-making, he said. However, imposing strict fairness criteria remains popular with both researchers and practitioners. "We believe that this fact highlights the need for a robust discussion about why those who advocate for the use of fairness constraints do so," Nyarko said.

"Do formal fairness criteria accurately capture people's views on what it means to make an ethical decision, and should therefore be incorporated? Is adherence to a 'fair' decision-making process—for

example one that does not use race—desirable for its own sake, or is it just a useful heuristic that often leads to more equitable outcomes? Only if we have clarity on these normative and ethical questions can we hope to make progress towards understanding what it means for algorithmic decisions to be fair. Hopefully, this will also lead to more homogeneity in approaches."

Establishing a framework around disparate debates

Nyarko stressed that numerous studies, especially in the medical context, have examined the impact of imposing fairness constraints like race- or gender-neutral decisions. The new paper is designed to "give a unifying framework to these discussions," he said. "You see a lot of individual papers, spread out across disciplines, that touch on issues of algorithmic fairness, but we think the debate needs structure and that's what we set out to accomplish," he said. "I think a lot of these individual discussions have not been well connected to broader ethical discussions of fairness."

The paper tackles each of the three most typical fairness constraints, all of which are "intuitively appealing," but which can lead to results that are bad for individuals and society as a whole, they write. The fairness constraints are:

1. Blinding, in which one limits the effects of demographic attributes—like race—on decisions
2. Equalizing decision rates across demographic groups (for example, requiring that the share of patients referred for further diabetes testing is the same for Asian and non-Asian patients)
3. Equalizing error rates across demographic groups (for example, requiring that the share of patients who are erroneously excluded from testing even though they have diabetes (so-called false negative rate) is the same for Asian and non-Asian patients)

The paper offers several recommendations to people training algorithms to assist in decision making, including that they understand the pitfalls of "label bias."

"There is a very widely held belief in the machine learning literature that giving more data to the algorithm can't do any harm," Nyarko said. "Either the information is helpful in making the prediction or it gets discarded. But this is only true if the thing we train the algorithm to predict is the thing we really care about. However, it turns out that these two routinely diverge.

"In the context of criminal justice, for example, a judge who is making a detention decision might want to know how likely it is that a defendant will recidivate. This will help the judge decide whether the defendant should remain in jail or can be released. Algorithms routinely assist judges in making these decisions. However, these algorithms have never been trained to predict the likelihood of recidivism. After all, whether someone commits a crime is not something that is really observable at scale.

"All we know, and all that an algorithm is trained to predict, is whether a defendant is likely to be rearrested. Whether someone is rearrested for a crime can depend, in large part, on whether there is a lot of police presence in the area where they live. This type of label bias is very common, and we show that it has important implications for how algorithms should be trained.

"For instance, in our example of recidivism risk prediction, we show that giving access to a defendant's ZIP code to commonly used algorithms improves their prediction for whether a defendant will be rearrested. However, due to the disparities in policing across neighborhoods, giving access to ZIP codes makes the same algorithms worse at predicting whether the defendant will recidivate. More generally, our findings call

into question the common wisdom that adding more data can't make our algorithmic decisions worse."

More information: Alex Chohlas-Wood et al, Designing equitable algorithms, *Nature Computational Science* (2023). DOI: 10.1038/s43588-023-00485-4

Provided by Stanford University