Beyond the Algorithm

Something interesting from Exhibit 1 was their finding that their algorithm appeared to work fairly across all three racial groups for the minimum-risk and high-risk categories, but there were still substantial differences at the middle risk levels. It would be a good idea to investigate why this is and how we might adjust the algorithm to fix it.

Exhibit 2 (pg 7) talks about an AUC statistic, which stands for Area Under Curve of an ROC curve. This is a great resource that explains ROC and AUC: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
In short, ROC deals with true and false positives, and the AUC can be interpreted as the probability that we have a true positive. So in this case, the AUC of .719 says that 71.9% of the algorithm's predictions were correct. In the context of the article, all three race's AUC's were above .7 and within a window of .016, which suggests that the algorithm was generally effective and equal for each group (At the high risk level).

Page 8 talks about a lot of what we have already been discussing, how one of the main sources of bias comes from the disproportionate rates at which minorities are arrested. This will in turn give them a higher probability of having a history of arrest, which then makes them look higher-risk to the algorithm.

One way of overcoming this bias would be to create separate algorithms depending on race, however as the article discussed this would be impossible ethically and legally, and generally would just be a step in the wrong direction. Something that we can do however is try to reduce the amount of weight the algorithm places on past criminal records, especially if the crimes are petty.

Exhibit 3 builds on this idea a little bit. They adjusted the algorithm to combine the two highest risk groups, and saw that Black people were much more likely to be classified in the upper risk groups. They note that this is not necessarily a problem with the algorithm, but "base rate" meaning that the data the algorithm is built on is already corrupted in that the police and judges are not always unbiased in who they are arresting and how they are processed.

Algorithmic Justice

The article begins by talking about algorithmic governance, explaining how big data, AI, and algorithms already automate many instances of decision making. In the justice sector, these cases of automated decision making can have implications not seen elsewhere. This leads into the "Existing research on algorithmic justice" section beginning on page 3.

One of the concerns deals with actuarial risk assessment. Actuarial risk assessment is a calculated prediction of that likelihood that an individual will pose a threat to others within a certain time period.

The article has a section called "the risks and pitfalls of algorithmic justice" beginning on pg. 7. A good point they make is the quality of the data depending on the domain these

algorithms are working in. Their example is that false-positives in earthquake prediction are more acceptable than those in predictive policing, where a false positive affects individual liberties. Criminal data is also never fully reported. This affects the accuracy of any algorithm since it technically cannot encompass everything. The data calculated, then needs to reflect the reality of the environment as closely as possible. An example used is the COMPAS algorithm, where there was a higher false positivity seen in black defendants compared to white ones, meaning blacks had a higher risk score, in general, yet they reoffended less.

The section, “Human or Machine Bias” warns against unintentionally using data that isn’t clean of social or cultural circumstances. Even in analyzing language, natural language contains human biases. Then it just comes down to deciding between wanting human bias or machine bias (page 11).

Using Algorithms to Address trade-offs

Talks about algorithms being used to predict violent reoffending. In a sample of about 68,000, the AUC was .71-.72, which is the probability of a true positive. These algorithms had different associations with race, from r=0 to r=0.21. R = 0 means weak or no association and .21 means generally weak association.

The trade-off mentioned in the article is one that justifies the disparities in COMPAS algorithm. On pg. 4 they argue that when there are different rates of reoffending across racial groups, the algorithm will classify certain groups as high risk more frequently as others. This is why it is argued that the results of the COMPAS algorithm given in the previous section are a result of an algorithm that is well-calibrated to the population.

A statistical method they used in page 5 is to see how alternative approaches affect the trade-off mentioned above. First there are race-included and race-omitted algorithms to serve as benchmarks. Then the other three include omitting different aspects to see if they eliminate bias.

<u>Methods used</u> :

- For PCRA algorithm(pg. 10)- Used AUC and correlation to black race to measure accuracy and bias
- Compared PPVs, FPRs, and FNRs across algorithms.
    - PPVs: the probability that a positive from the algorithm is a true positive
    - FPRs, FNRs: False positive and false negative rates