RISK ASSESSMENTS AND RACIAL BIAS

Presentation by Megan Stevenson, Assistant Prof. of Law at George Mason University

Goals for today:

- Clarify what people mean when they call risk assessment tools racially biased
- Evaluate commonly suggested potential remedies
- Suggest practical rules-of-thumb for promoting racial fairness with algorithms
  - This involves thinking beyond the algorithm itself to questions of how it is developed, used, and monitored.

RACIALLY BIASED ALGORITHMS

What do people mean by that?

Why do people think risk assessments are racially biased?

- **Disparate impact on African-Americans**
  - Black defendants usually have higher average risk scores
  - Long history of racial injustice in US
  - Concerns that risk assessment will entrench or exacerbate race disparities

- **Disparate impact is not always unfair.**
  - But it can be unfair if it imposes unwarranted and disproportionate harms on a disadvantaged group
  - Important: who gets to decide what is “unwarranted” or “disproportionate”??

Why do people think risk assessments are racially biased?

- **Because they are trained on real world data**
  - **The real world is racially biased**
  - **This is a fundamental reason why no risk assessment can truly be "race neutral"**
- Our proxies for offending rates are racially biased
  - We can't see true crime rates, we see arrest rates, conviction rates, etc.
    - Minority neighborhoods more heavily policed
    - Wealthier people can afford fancy lawyers
    - Racial bias/stereotypes may affect charging decisions, etc.
- Even if we had perfect data about offending rates, that doesn't mean the data is unbiased
  - Someone may commit a crime because they are poor
  - But they may be poor because of a legacy of racial oppression

PROPOSED REMEDIES

Evaluating commonly proposed remedies for racial unfairness in algorithms

“De-biasing” the data

- Could you simply correct the data to account for the fact that arrest is a racially biased measure of offending?
  - We know that despite similar levels of drug use, black people are more than twice as likely to be arrested
- Fundamental problem: we don’t have any data on actual illegal behavior
  - We only have racially filtered proxies for crime, like arrest
  - Drug use might be the one exception
- How much “de-biasing” to do?
  - Extremely speculative
  - Subject to legal challenge
  - Not currently workable

Banning the use of “race-proxy” inputs

- What does “race-proxy” mean?
  - Not really defined
- Many inputs to a risk assessment are correlated with race
  - Socio-economic markers such as employment, housing status
  - Criminal history variables such as # of prior arrests, convictions
  - ZIP code
- A “race-proxy” could mean
  - So strongly correlated with race that it is a “stand-in” for race
  - A factor that is more correlated with race than it is with recidivism
  - People generally think of it as a sneaky way of getting race into a risk assessment

Banning the use of “race-proxy” inputs

- Usually what people mean by this is to **ban non-culpable inputs** that are correlated with race
  - Employment
  - Housing
  - Education
  - Marital status
  - ZIP code
- This doesn’t make a risk instrument “race neutral” but is probably still a good idea
  - Doesn’t reduce accuracy much
  - Reduces potential legal challenges
  - May reduce the race gap in risk assessments (unclear)
  - Will reduce public distrust of risk tools (perceptions of legitimacy)

Selecting target variables that are less racially biased

- Target variable: the outcome that the tool is trained to predict
- Arrest for serious violent crime may be less racially biased than arrest for lower-level offenses
  - Why?
    - Policing resources are limited; arrests for low-level crimes more likely in heavily policed neighborhoods (where minorities live)
    - More discretion in arrest for low-level crimes
- What happens if you train an algorithm to predict arrest for serious violent offense?
  - Likely the risk scores will become more racially disparate
  - The people who get arrested for serious violent offenses are disproportionately young black men
  - Tool will be inaccurate – harder to predict low frequency events

Equalizing false positive rates

- **False-positive rate**: the fraction of people labeled high risk who do not reoffend
- Equalizing false-positive rates would involve making a risk tool that was differently calibrated for each race*
  - Affirmative action

- **This is a bad idea**
  - Politically unfeasible
  - Probably unconstitutional
  - Resulting risk instrument would be hard to use/interpret

* Assuming that the reoffending measure is different across races, which it almost always is.

Demonstrating predictive parity

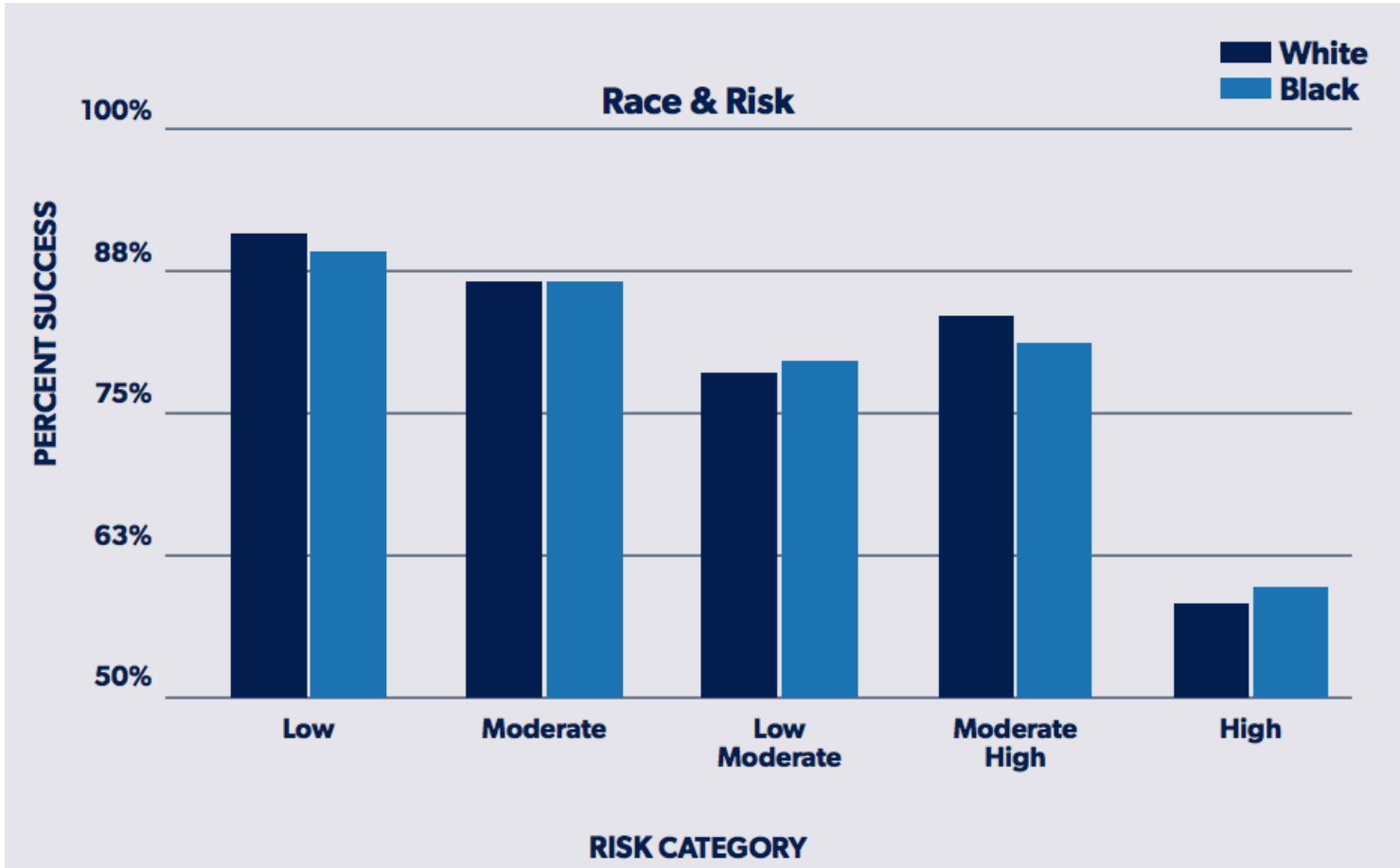

FIGURE 1: THE PUBLIC SAFETY ASSESSMENT – COURT ANALYSIS OF RACE AND GENDER

From “Pretrial Risk Assessment Can Produce Race-Neutral Results” –PJI, 2017

Demonstrating predictive parity

- In the previous slide, pretrial success is defined as:
  - Making all court appearances
  - Not being rearrested during the pretrial period

- What determines pretrial success?
  - A person's propensity to commit crime or FTA: their "risk"
  - The conditions that were imposed on them by the court
  - Their propensity to be arrested conditional on committing a crime
  - The propensity of the courts to record an FTA for someone who is not there on their court date
- Lots of things beyond a person's innate "risk" can influence pretrial success!

Abandoning algorithms altogether

- No easy fixes
  - Should we abandon algorithms and stick with human intuition?
- All of the concerns we have listed with racial fairness and algorithms apply equally to human predictions!
  - The disparate false positive rates issue raised by ProPublica applies equally to human predictions
  - Humans also make predictions based on racially disparate real-world data (their experiences)
- Plus…
  - Human predictions are “black-box”
  - Humans may be explicitly racist
  - You have no control over what factors enter the human predictions

To summarize:

- De-biasing the data is not possible
- Avoiding certain “race-proxies” may increase perceived legitimacy but will not alleviate fundamental issues of racial fairness
- Choosing a less-biased target variable could actually increase racial disparities
- Equalizing disparate false positive rates is not feasible
- Demonstrating predictive-parity is not enough
- Avoiding algorithms altogether is not a solution

SYSTEM AND PROCEDURAL FAIRNESS

Looking outside the algorithm to increase fairness in the use of risk assessment

Beyond-the-algorithm suggestions for improving fairness

- **Use risk assessments to increase release without onerous conditions**
  - Risk assessments less likely to be perceived of as racist if their use results in lower incarceration for people of color
    - Use the “high risk” label sparingly
    - Ensure due process for defendants before detention or electronic monitoring (not just triggered by high-risk label)
    - Make pretrial release without onerous conditions the presumptive default
    - Monitor use of risk assessment to increase accountability

Beyond-the-algorithm suggestions for improving fairness

- **Evaluate the impacts of adopting risk assessment**
  - Make data open and accessible to third party researchers and the communities most affected
  - Check to see:
    - If risk assessment is being used to reduce racial disparities
    - If it's resulting in lower detention rates
  - Adjust and re-evaluate if goals are not being met

Beyond-the-algorithm suggestions for improving fairness

- **Create accountability measures for those tasked with using the tools**

- People respond to what's being measured, so monitor and evaluate
  - Do black defendants with a "low risk" label get released at the same rate as white defendants with the "low risk" label?
  - How often are the "presumptive defaults" being overruled?
  - Etc.

- Integrate this into performance evaluations

Beyond-the-algorithm suggestions for improving fairness

- **Include diverse stakeholders in each step of the process:**
  - Designing or choosing the risk assessment tool
  - Designing the implementation schema
  - Monitoring and evaluating the impacts of the tool
- **Make the process transparent**
- **Train users about the potential sources of racial bias in the tools**