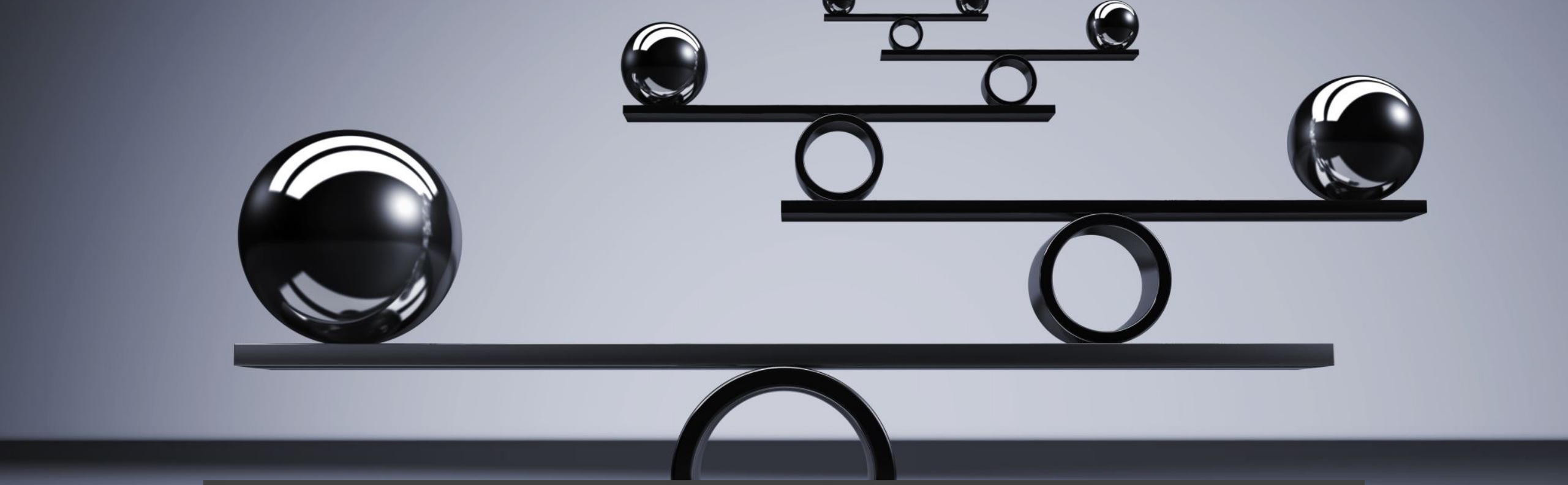
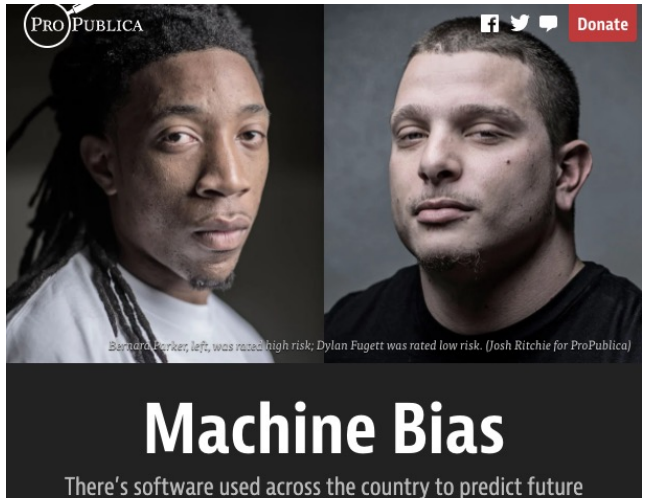


ALGORITHMS AND FAIRNESS

Aksel B. Sterri, Embedded EthiCS
Computer Science 1, Harvard University

UNFAIR ALGORITHMS?

- Algorithms are increasingly used in *screening decisions* like hiring, university admissions, sentencing.
- Biased and unfair?
- Were implemented partly to make fairer and more accurate decisions!
- What should we believe? Can both claims be true?



OCTOBER 11, 2018 / 1:04 AM / UPDATED 2 YEARS AGO

azon scraps secret AI recruiting showed bias against women

Dastin 8 MIN READ

RANCISCO (Reuters) - Amazon.com Inc's AMZN.O e-learning specialists uncovered a big problem: their ng engine did not like women.

QUESTION

- Why would algorithms help us make fairer decisions?
- Why would they make unfair decisions?
- 1-2 points on each.
- What's your overall assessment? Algorithms source for good or bad?

WHAT STANDARD?

AIM

1. Examine flaws in human decision-making
2. Learn to distinguish between noise, bias, and fairness
3. The role algorithms play in reducing noise and bias (but also perpetuating it)
4. Reflect on whether eliminating noise and bias for an algorithm to be fair.

HUMAN DECISION-MAKING

EVIDENCE #1

Harsher sentences if the local football team loses.
According to a study of 1.5 million judicial decisions over three decades.

Lenient sentences on people's birthday.
According to a study of six million decisions made by judges in France over twelve years.

#EVIDENCE 2

Study of 208 federal judges considering16 hypothetical cases:

1) Should someone go to jail and if so, 2) for how long?

1) Unanimous agreement only in 3 of 16 cases.

2) Substantial variation. In one fraud case: 8.5 years mean prison term; the longest was life in prison.

EVIDENCE #3

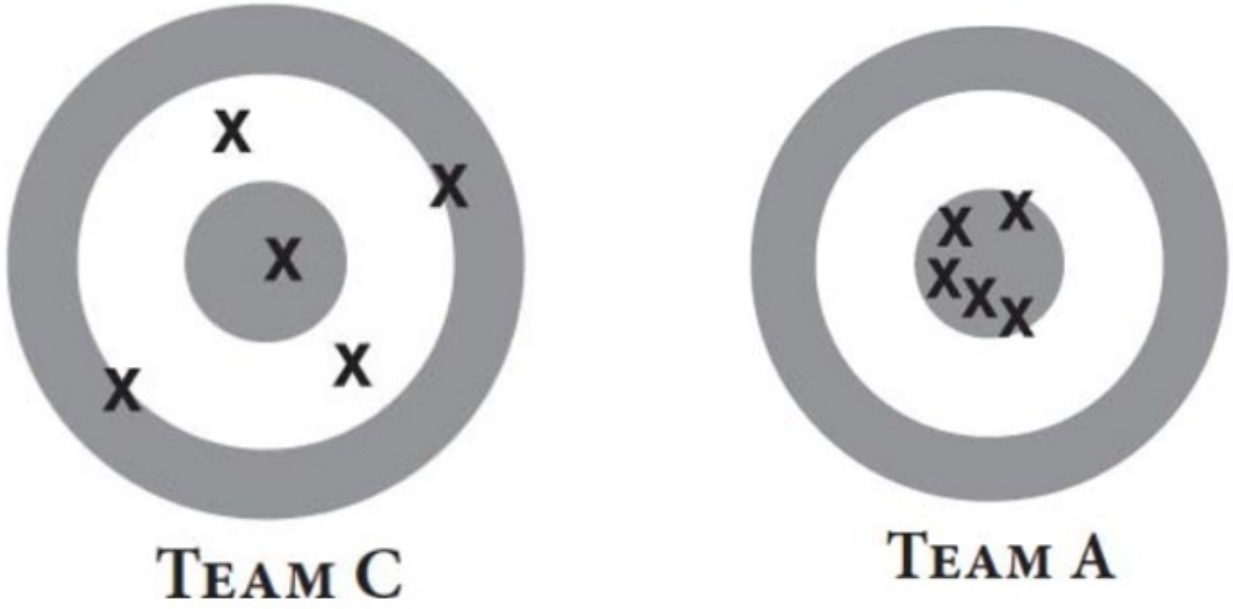
Fictitious CVs sent to employers in Boston and Chicago. **Identical** except half white-sounding name and half African-American- sounding name.

CVs with white names received 50% more call-backs from employers

Doctors shown two equivalent patient histories:

The chances of recommending a beneficial procedure were 40 percent lower for women and minorities than white males.

CV study: Bertrand and Mullainathan 2004
Doctor study: (Schulman et al. 1999).

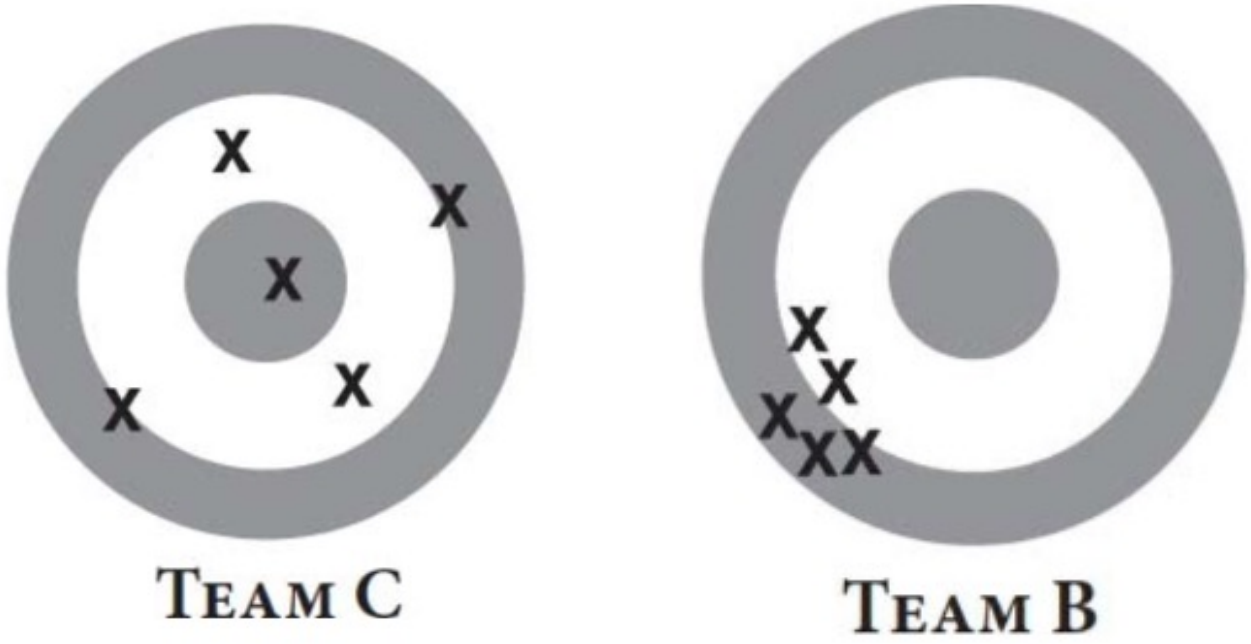


NOISE: UNSYSTEMATIC INACCURACY

THE DECISION-MAKER MATTERS

1) Unanimous agreement only in 3 of 16 cases.

2) Substantial variation. In one fraud case: 8.5 years mean prison term; the longest was life in prison.



BIAS: SYSTEMATIC INACCURACY

IRRELEVANT FACTORS MATTER

Harsher sentences if the local football team loses.

Lenient sentences on people's birthday.

ONE'S GROUP MATTERS

CVs with white names received 50% more call-backs from employers

The chances of recommending a beneficial procedure were 40 percent lower for women and minorities than white males.

CV study: Bertrand and Mullainathan 2004
Doctor study: (Schulman et al. 1999).

VIOLATES A MINIMAL ACCOUNT OF FAIRNESS

1. Irrelevant factors should not influence the decision.
2. The identity of the person who assesses the case should not impact the decision.

WHAT GOES WRONG?

1. Sometimes we are straightforwardly racist and sexist.
2. Other times the decision is too complex.
    1. We do not always know what we are aiming for
    2. Predicting who will do well at a job requires knowing which factors predict performance.
    3. We, therefore, rely on stereotypes and imprecise heuristics.

COULD ALGORITHMS HELP?

REMEMBER THE PROBLEM

- The problem is to figure out who is guilty, who is at risk of reoffending, who should be hired, and who should be admitted to a university.
- Requires knowledge about what we are trying to achieve (the objective)
- And what factors predict performance pertaining to those objectives.

CREATING A SCREENING ALGORITHM

STEP 1: COLLECT OBSERVATIONS ON PAST CASES

STEP 2: CHOOSE OBJECTIVES

STEP 3: FIND PREDICTORS

USING THE SCREENING ALGORITHM

1

STEP 1: COLLECT OBSERVATIONS ON CANDIDATES

2

STEP 2: PREDICT SUCCESS FOR CANDIDATES

3

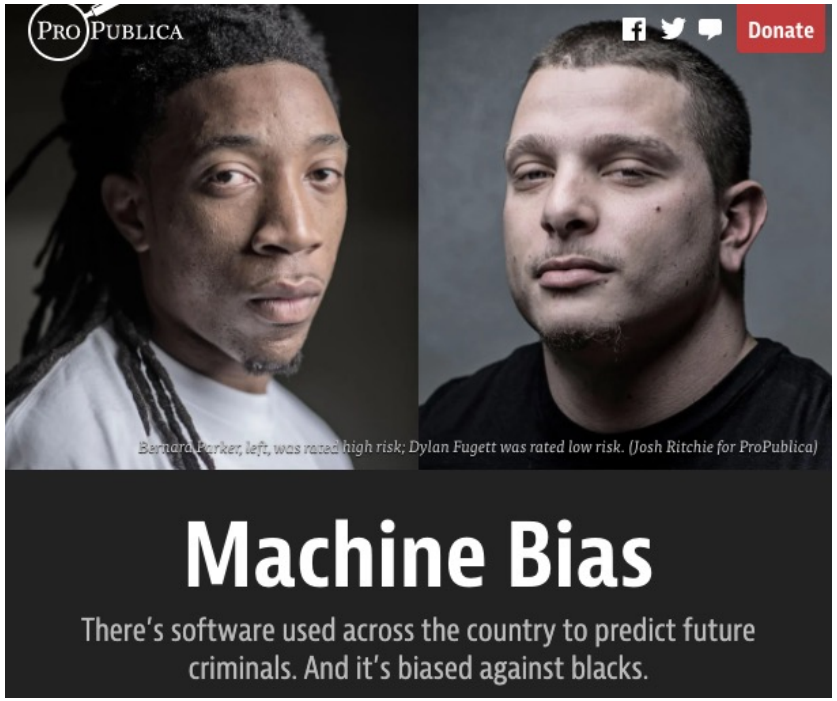
STEP 3: RANK ALL CANDIDATES

SHOULD REDUCE BOTH NOISE AND BIAS

1. It forces us to be explicit about our aims.
2. It does not rely on stereotypes about what factors contribute to success but on observed predictors in past datasets.
3. It applies the same algorithm (logic) to every case. It therefore reduces noise.

EXAMPLE

- Judges in New York must make decisions of whether to release a criminal defendant pre-trial based on a predictions of risk of failure to appear in court (FTA).
- Kleinberg et al. (2018) built a machine-learning algorithm to predict risk.
- Finding: Judges "detain many low-risk people and release many high-risk ones."
- Use of algorithm: "could reduce the jail population by 42 percent without increasing FTA rates at all" [because of increased accuracy.]



RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

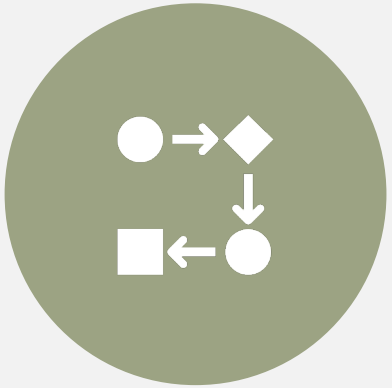
SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

BUT IF THEY ARE SO GOOD, WHAT GOES WRONG AT GOOGLE AND WITH COMPAS?

WHERE CAN NOISE AND BIAS CREEP IN?

STEP 1: COLLECT OBSERVATIONS ON PAST CASES

STEP 2: CHOOSE OBJECTIVES



STEP 3: WHAT CHARACTERISTICS PREDICT "PERFORMANCE" (BY HUMAN OR MACHINE)

SOURCES OF BIAS

- *Wrong objective:*
  - Trained on previous hires (like Google's algorithm) or on who are liked by their bosses or co-workers
  - Risk of reoffending: depends on who the police are after.
  - Perhaps preferable: objective criteria of success.
- *Non-representative training data*:
  - Too little data for one group (makes it noisier for that group)
  - Skewed data on one group
  - Hide race, gender, and other characteristics (may paradoxically increase bias)
  - Correlation is not causation. Hostile workplace (sexual harassment) and worker performance
- *Wrong inference*:
  - A hostile workplace calls for measures to improve workplace culture, not avoid hiring the oppressed group.

SUPPOSE WE COULD FIX THESE PROBLEMS

- We choose an appropriate objective (one's contribution to profit)
- Representative training data
- But the result is that the algorithm still ranks Blacks disproportionally low for a job.

- What then? Is the algorithm biased? Is it unfair? Discuss.

OTHER SOURCES OF UNFAIRNESS

- If profit is the objective, that could also create unfairness if customers are racist.
- What about historical injustices?
- Should there be a role for affirmative action?

A BROADER ACCOUNT OF FAIRNESS

1. Irrelevant factors should not influence the decision.
2. The identity of the person who assesses the case should not impact the decision.
3. **Does the decision contribute to "societal fairness", a fairer overall distribution of benefits and burdens?**

SURVEY

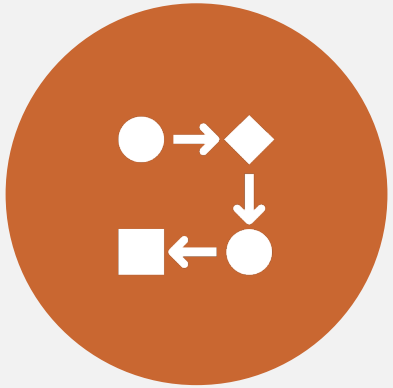
ALGORITHMS CAN ALSO BE USED TO REDUCE BIAS

- Human decision-making black box.
- Algorithms can be analyzed and interrogated.
- Why do they make the decisions they make?

IDENTIFY BIAS

STEP 1: WHAT CANDIDATES AND WHAT CHARACTERISTICS?

STEP 2: WHAT OBJECTIVES?



STEP 3: FIND CORRELATIONS