UNFAIRNESS IN ML ALGORITHMS

Sophie Gibert, 20 October 2021

Embedded EthiCS @ Harvard

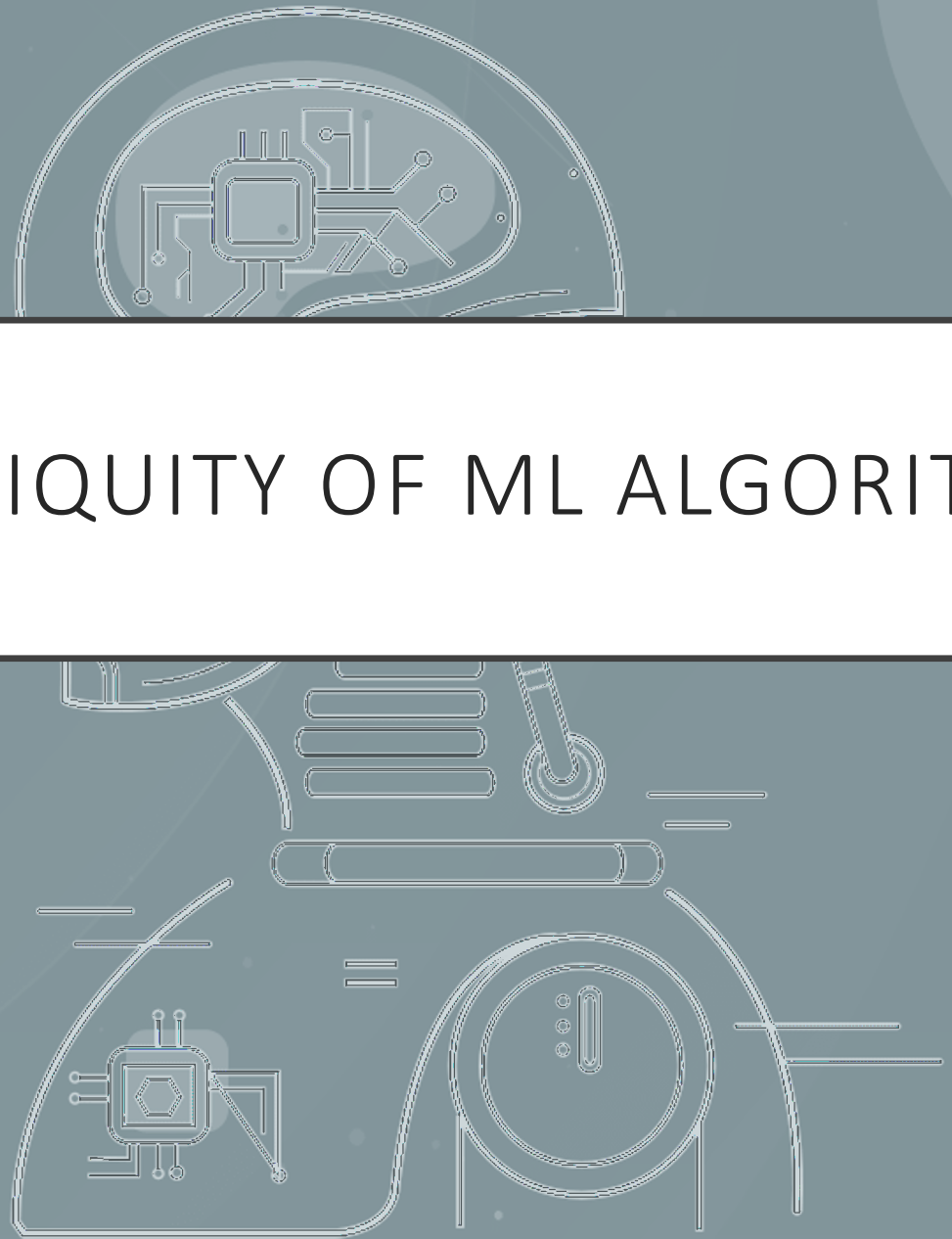
WHO AM I?

Sophie Gibert, PhD Candidate in MIT's Philosophy Department

Graduate Fellow for Embedded EthiCS @ Harvard

Research interests: Ethics, philosophy of action, bioethics

THE UBIQUITY OF ML ALGORITHMS



WHY USE ML ALGORITHMS?

- Decisions, categorizations, and predictions profoundly affect our lives
- We want them to be consistent and evidence-based, and to consider relevant factors
- Humans are limited; algorithms are often better at achieving these goals

THE PROBLEM OF UNFAIR BIAS

ML algorithms exhibit biases, and these biases often seem unfair

Examples:

- Amazon hiring
- PredPol predictive policing
- Skin-cancer detection

TODAY'S PLAN

1. How bias enters ML algorithms
2. Case study in healthcare
3. What we can do

HOW DOES BIAS ARISE?

A CASE STUDY IN HEALTHCARE

BACKGROUND

- Across the US, health systems and insurers use algorithms to identify patients with complex health needs
- These patients receive extra healthcare resources
- The goal is to improve outcomes and reduce costs

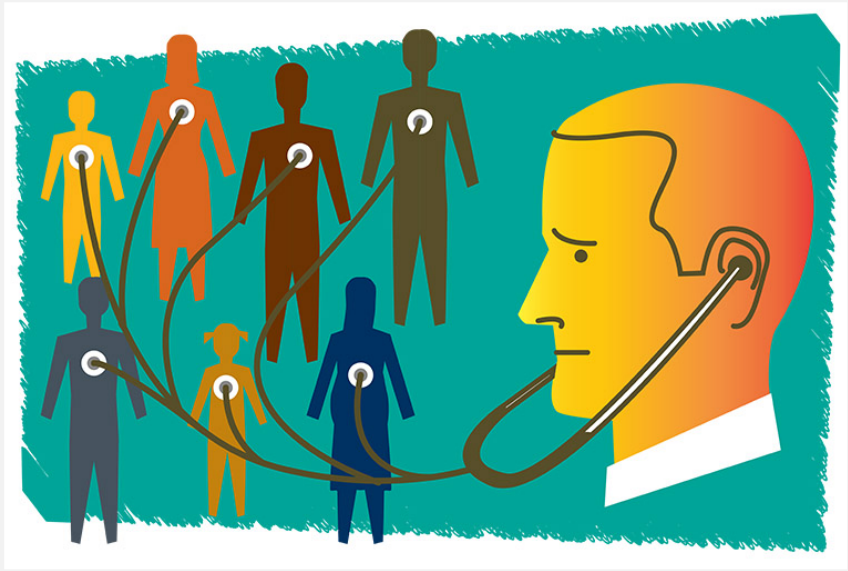
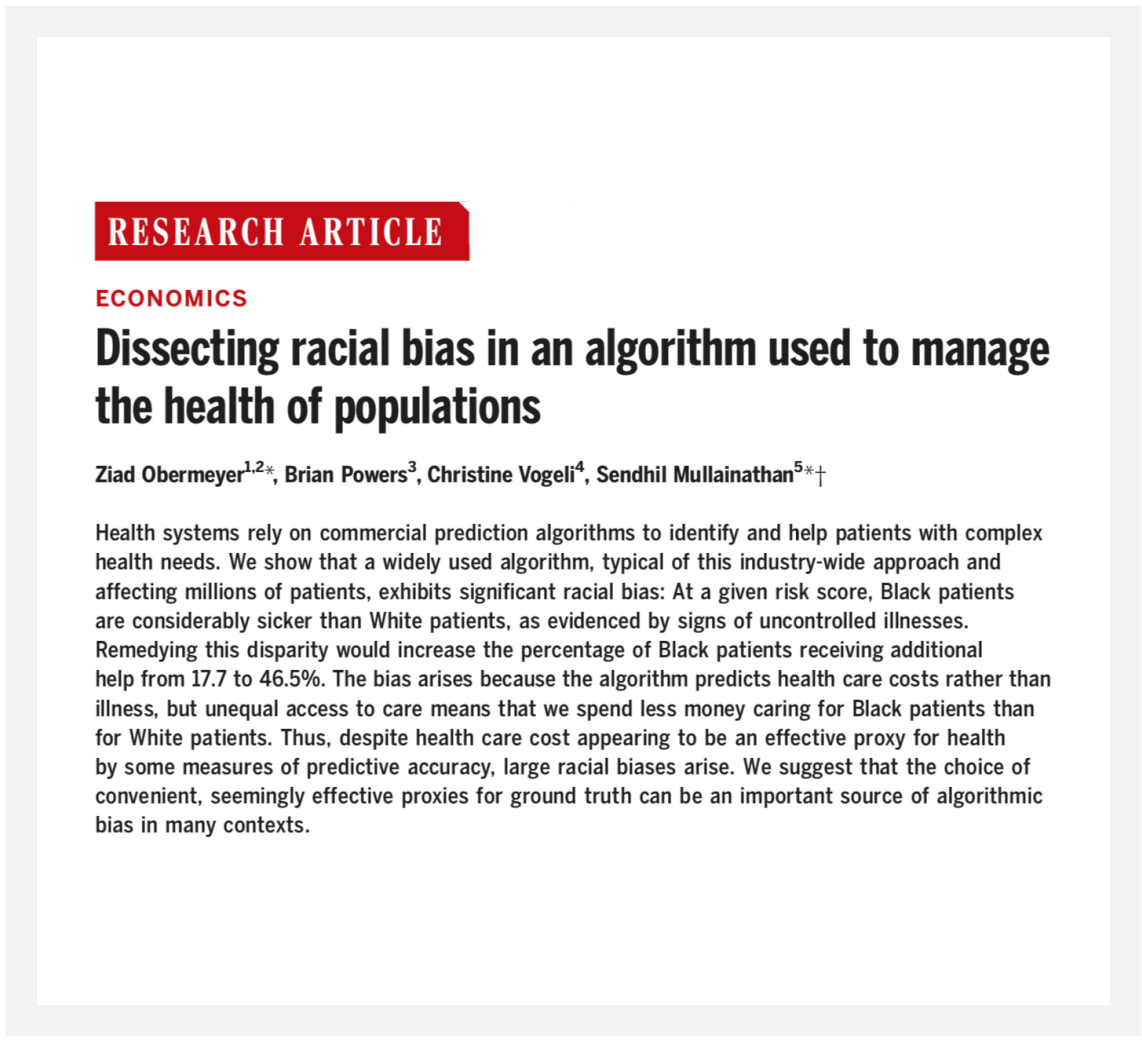


Illustration by Hulda Nelson

BACKGROUND

- October 2019: A major healthcare risk prediction algorithm was shown to be racially biased
- Algorithm assigned risk scores to patients; those in the 97th percentile or above were automatically enrolled in the high-risk program
- Designers chose to predict healthcare costs as a proxy for need



RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

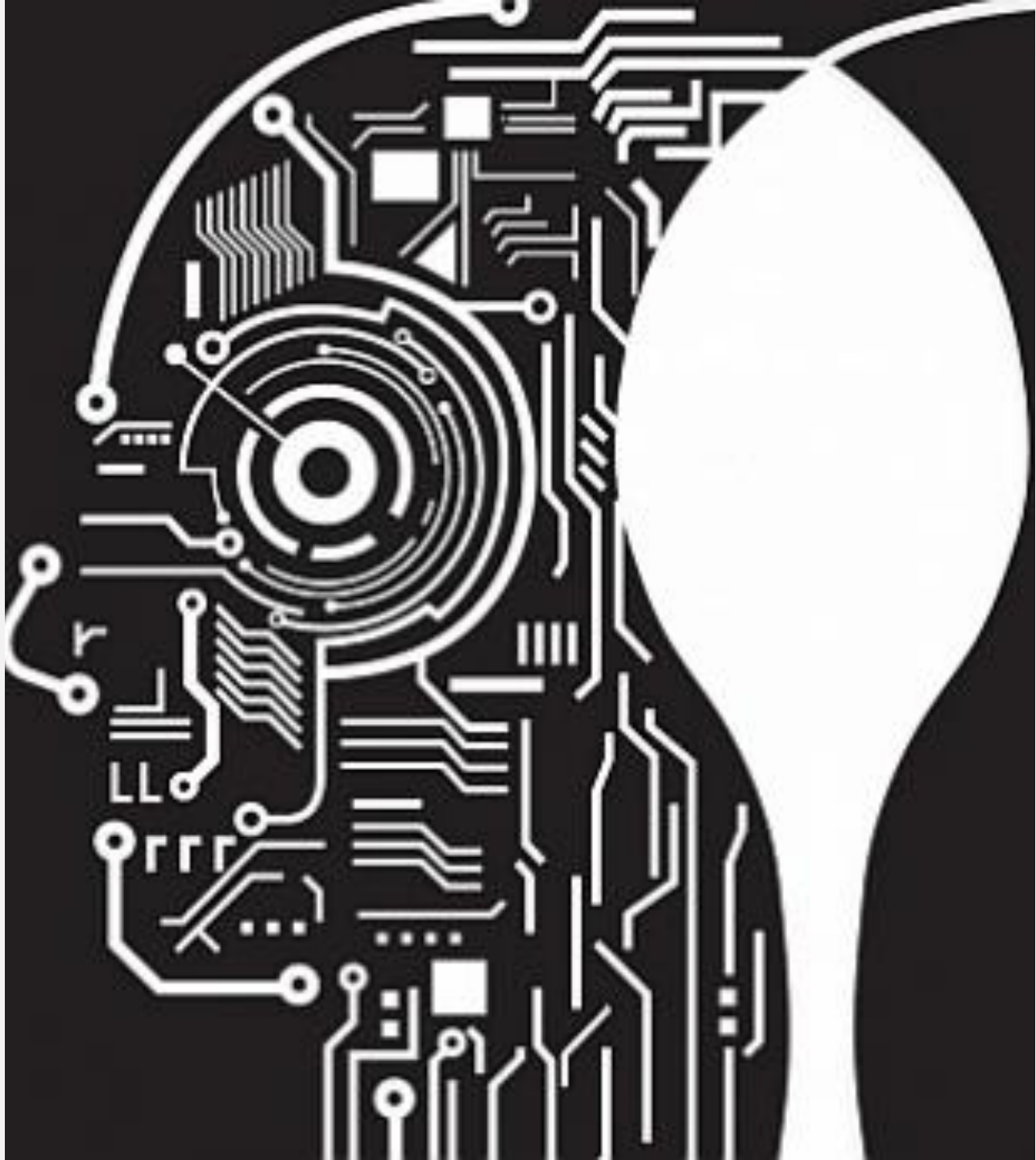
DISADVANTAGE TO BLACK PATIENTS

Illustration by Erin Robinson

- **Well-calibrated for costs across racial groups**: At each risk level, Black and White patients had similar costs in the next year
- **Not well-calibrated for health outcomes across race**: At each risk level, Black patients were sicker

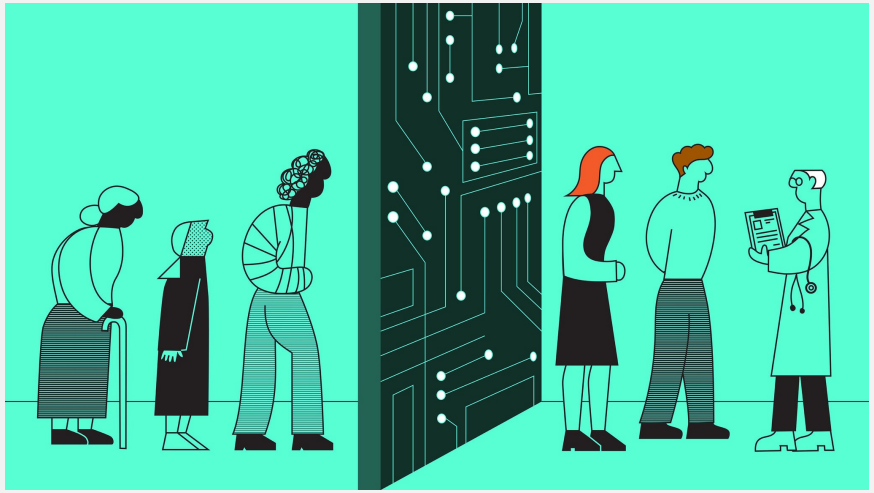
DISADVANTAGE TO BLACK PATIENTS

- Among those classified as high-risk, Black patients had 26.3% more chronic illnesses than White patients despite receiving similar risk scores



EXPLAINING THE DISPARITY

- The algorithm predicts **healthcare costs**
- Race is *not* a predictor; but some proxy variables for race are
- At any given level of health, Black patients generate lower healthcare costs than White patients
- Two mechanisms:
  - Poverty and barriers to access
  - Low uptake due to bias and lack of trust



DISCUSS: (WHY) IS IT UNFAIR?

1. Do you agree that it seems unfair for Black patients to receive lower risk scores than their White counterparts in these circumstance?
2. Why might someone think this *is* unfair?
3. Why might someone think this is *not* unfair?

A LUCK EGALITARIAN ANSWER

- According to **luck egalitarians**, inequality is fair when, and only when, it results from *free choice*, as opposed to luck
- Perhaps the algorithm is unfair because it includes predictors that are not entirely a matter of free choice

Illustration by Tim Cook

TOO QUICK?

- For centuries, luck egalitarians have been vexed by the following question: Which inequalities are chosen, and which are the result of luck?
- Critics argue that some inequalities should be compensated even if they are the product of free choice
  - Socially important roles
  - Goods that promote democracy, solidarity, and respect (E.g., voting rights, healthcare?)

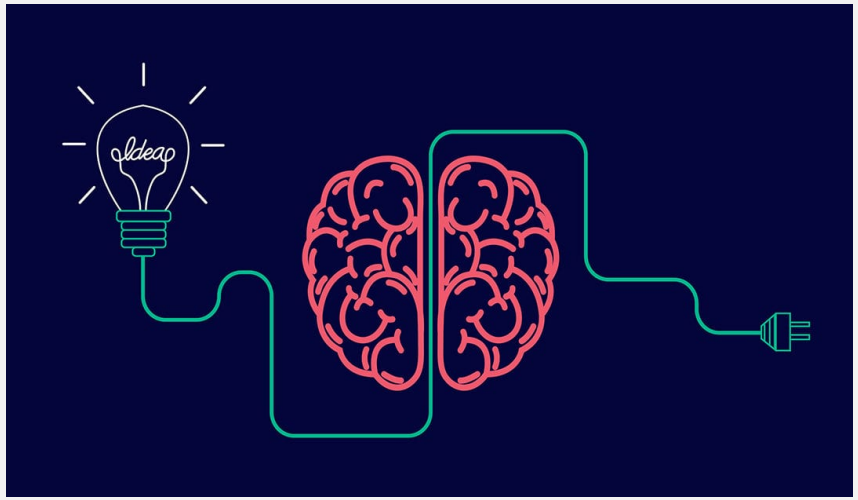
Illustration by Monica Garwood

RELABELING AS A SOLUTION

Relabel: Predict something else

Some options:

- Catastrophic healthcare utilization
- Number of chronic conditions
- Number of illnesses



Predicting an index variable that combines health and costs reduces bias

WHAT CAN WE DO?

ASK QUESTIONS

Problem formulation

- Is a proprietary ML algorithm an ethical solution to the problem? Is this a decision for which the decision procedure should be transparent?
- Is it fair for this outcome to depend on this label?

ASK QUESTIONS

Dataset construction

- Is the training dataset representative of the group on which the algorithm will be deployed?
- Do the training data reflect any unfair biases that will be reproduced by the algorithm?
- Is it fair for this outcome to depend on these predictors, given that such biases will be reproduced?

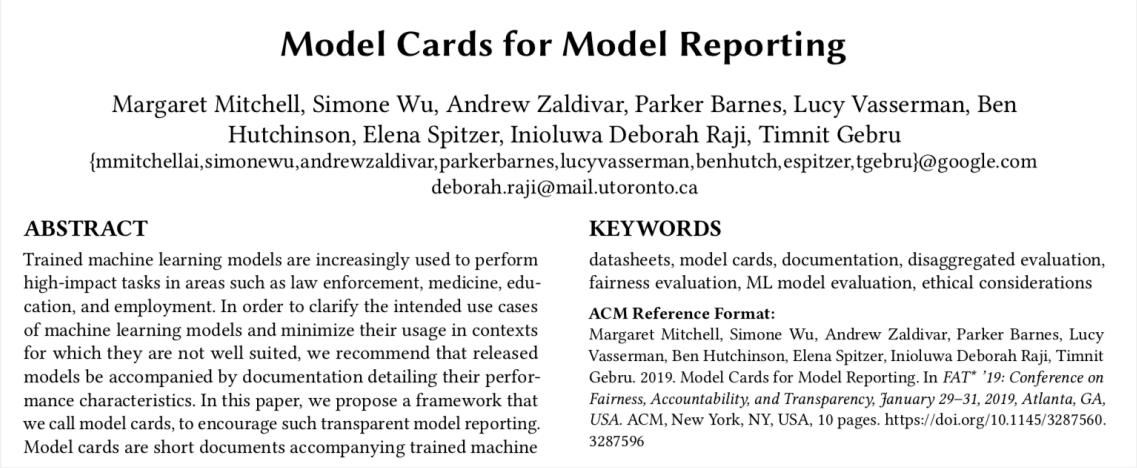
ASK QUESTIONS

Deployment

- Is the algorithm being deployed in a group that was adequately represented in the training data?
- What do users of the algorithm need to know to use it appropriately?

MODEL CARDS FOR MODEL REPORTING

- Provide information that helps people decide when and how to use a given ML algorithm
- Explain how the model was built, what assumptions were made during development, how it performs in different cultural, demographic, and phenotypic groups
- Make ethical recommendations



Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

THANK YOU

Contact: Sophie Gibert,
sgibert@g.harvard.edu

Survey:
https://tinyurl.com/CS109aF21

REFERENCES

Anderson, Elizabeth. 1999. “What is the Point of Equality? *Ethics* 109 (2): 287-337.

Appiah, Kwame Anthony. 2020. “The Case for Capitalizing the ‘B’ in ‘Black.’” *The Atlantic*.

Binns, Reuben. 2018. “Fairness in Machine Learning: Lessons from Political Philosophy.” *Proceedings of Machine Learning Research* 81: 1-11.

Corbett-Davies, Sam et al. 2016. “A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It’s Actually Not That Clear.” *The Washington Post*.

Larson, Jeff et al. 2016. “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*.

Mitchell, Margaret et al. 2019. “Model Cards for Model Reporting.” *Association for Computing Machinery*.

Obermeyer, Ziad et al. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366: 447-453.